Modeling the Energy Efficiency of GEMM using Optical Random Access Memory

Bingyi Zhang*, Akhilesh Jaiswal†, Clynn Mathew†, Ravi Teja Lakkireddy†, Ajey P. Jacob†,
Sasindu Wijeratne*, Viktor Prasanna*
*Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA
†Information Sciences Institute (ISI), University of Southern California, Marina Del Rey, USA
Email: bingyizh@usc.edu, {akjaiswal, cmathew, lakkired, ajey}@isi.edu, {kangaram, prasanna}@usc.edu

***Abstract*—General matrix-matrix multiplication (GEMM) is the key computation kernel in many applications. GEMM has been supported on various hardware platforms, including CPU, GPU, FPGA. To optimize the performance of GEMM, developers use on-chip electrical static random access memory (E-SRAM) to exploit the data locality of GEMM. However, intensively accessing E-SRAM for GEMM can lead to significant energy consumption, which is not energy-efficient for commercial data centers.**

In this paper, we evaluate the optical static random access memory (O-SRAM) for GEMM. O-SRAM is a promising technology that has extremely low access latency and low energy consumption compared with the traditional E-SRAM. First, we propose an O-SRAM based wafer-scale system for GEMM and a baseline E-SRAM based system. Second, we build the theoretical performance models of the two systems to analyze their energy consumption of on-chip memory accesses. Then, we conduct simulation-based experiments to evaluate the energy consumption of the two system. The evaluation results show that O-SRAM based system is 7× more energy efficient than the baseline E-SRAM based system.

***Index Terms*—Optical static random access memory, general matrix-matrix multiplication (GEMM), energy efficiency**

I. Introduction

General matrix-matrix multiplication (GEMM) is a key computation kernel in a broad range of applications, such as scientific computing [1], machine learning [2], [3], [4], etc. For example, in the well-known Convolutional Neural Networks (CNNs), the convolution operation on 2-D images can be transformed into GEMM operations [5]. State-of-the-art machine learning frameworks (e.g., Tensorflow [6], Pytorch [7]) regard GEMM as the key computation kernel to be supported. Many state-of-the-art computing libraries have supported GEMM on various computing platforms, including CPU, GPU, and FPGA. For example, Intel Math Kernel Library (MKL) [8] supports GEMM on CPU platforms. Nvidia CUDA [9] library supports GEMM on Nvidia GPU platforms. AMD Xilinx [10] developed libraries to support GEMM on various FPGA platforms.

To optimize the performance of GEMM, a commonly used strategy is to exploit the data locality of GEMM by using on-chip memory to cache the input matrices. Data caching in the on-chip memory can reduce the external memory accesses to the DRAM. For example, the GEMM function in Intel MKL library stores the matrix in the caches (L1/L2/L3 caches) of CPU. The GEMM function in CUDA library uses the cache of the GPU streaming processor to store the input matrices, that can increase the on-chip data reuse. The GEMM on FPGA exploit the on-chip block memory (BRAM, URAM) to cache the input matrix for data reuse [11]. While the data caching in the on-chip memory can reduce the external memory accesses for GEMM, there are still significant amount of on-chip memory accesses. As shown in [12], accessing the on-chip memory takes a significant amount (30%) of energy consumption for executing GEMM, becoming a significant energy bottleneck for data center. For example, Google's data centers [13] consumed around 15.5 terawatt-hours in the year of 2020. Therefore, reducing the energy consumption of accessing on-chip memory can potentially save the carbon emission by the data centers.

Optical random access memory (O-SRAM) has been looked upon as a promising pathway towards achieving ultra-fast and energy-efficient memory access [14]. However, despite several proposals [15], [16], [17], [18], [19], [20], [21], [22], [23], a robust, manufacturing-friendly, low-power O-SRAM had remained elusive. Recently, an ultra-fast and energy-efficient O-SRAM has been proposed featuring compatibility to existing silicon photonics foundry process [24]. The work in [24] has shown that an-optimized optical memory build using well-known silicon photonic device primitives can operate at the speed of 20 Gb/s and requires ultra-low static/switching energy consumption. Such recent advances in O-SRAM and their potential for mass manufacturing, makes O-SRAM as a promising alternative for traditional electrical SRAM for GEMM to save energy consumption. It is to be noted, however, that in general O-SRAMs have the disadvantage of high-area consumption. Specifically, despite achieving lower area compared to previous works, the O-SRAM presented in [24] is nearly 1000× larger than a traditional E-SRAM, which dramatically limits the density of O-SRAM. Therefore, it is non-trivial to evaluate energy efficiency of GEMM by simply replacing the E-SRAM with O-SRAM in a hardware platform (CPU, GPU, FPGA).

In this paper, we evaluate the energy efficiency of GEMM on a wafer-scale system with O-SRAM. Wafer-scale system [25] is a type of very-large integrated circuit built on an entire silicon wafer. For example, Cerebras [26], [27] has developed wafer-scale system to accelerate various applications of Artificial Intelligence. O-SRAMs are well-suited for wafer-scale systems since optical data can be seamlessly transferred across large-distances on a wafer-scale system with high-fidelity and at ultra-high speeds that are orders of magnitude faster than their electrical counterparts. Further, the large size of wafer-scale chips allow accommodating reasonable size of O-SRAM on-chip to accelerate GEMM. Thereby, we first propose an O-SRAM based wafer-scale system for GEMM to evaluate its energy consumption of on-chip memory accesses. At the same time, we build a baseline system with E-SRAM based on-chip memory. To evaluate their energy consumption, we build an accurate energy consumption model. Thereby, we perform simulation-based experiment to evaluate two systems. Our main contributions are:

- We propose a theoretical E-SRAM/O-SRAM based architectures for GEMM on wafer-scale systems.
- We build accurate energy models to estimate the energy efficiency of E-SRAM based system and O-SRAM based system for GEMM.
- We perform detailed evaluation to compare the energy efficiency of two systems. The experimental results show that O-SRAM based system is up to 7× more energy-efficient than the E-SRAM based system.

II. Background

A. Optical Random Access Memory

The optical memory used in our evaluation framework is shown in Feature 1. Photodiodes $D1$ and $D2$ along with ring resonators $R1$ and $R2$ form a cross-coupled pair, such that $R1$ in resonance ensures $R2$ is not in resonant to the incoming light wavelength and *vice-versa*. This creates a bistable circuit that can hold 1 bit of data in the optical domain. Photodiodes $D3$-$D4$ and ring resonators $R3$-$R6$ form the read-write path for the bistable optical circuit. The key highlights of the optical memory, shown in Figure 1 are as follows 1) use of well-known silicon photonic devices (photodiodes and ring resonator) make the proposed O-SRAM amenable to large-scale manufacturing on existing foundry process 2) the photodiodes as well as ring resonators are in reverse bias mode, thereby consuming minimal electrical power 3) functionally the O-SRAM is similar to E-SRAM featuring differential read and write operations, ensuring high robustness of optical data written/retrieved from the O-SRAM. A detailed description of the functioning and design of the O-SRAM can be found in [24].

B. Wafer-Scale system

Here, we will give a general introduction to the envisioned wafer-scale system based on O-SRAM. Arrays of O-SRAM cells would be fabricated on the wafer-scale chip in a typical silicon photonics foundry. The on-wafer optical memory would interface with electrical processing engines based on CMOS transistor technology. An electro-optic differential sense-amplifier would enable high-speed, low-energy conversion of optical data to electrical domain to be used by electrical processing engines. Thus, optical memory would serve as ultra-fast on-chip memory seamlessly integrated with electrical processing engines.

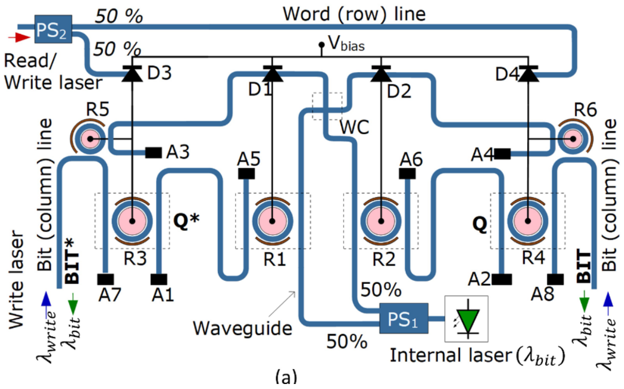

Fig. 1: Optical SRAM built using silicon photodiodes and ring resonators

III. Approach

A. Block Matrix Multiplication

We use $\boldsymbol{C} = \boldsymbol{A} \times \boldsymbol{B}$ to denote the GEMM operation, where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are dense matrices. For simplicity, we let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n \times n}$ where n is the dimension of the matrices. In the real-world applications, the matrices may not be fully stored in the registers of the processors. On-chip memory are used to store the input matrix for on-chip data reuse. To exploit the data locality of GEMM, the block matrix multiplication are used to execute the GEMM on the modern processors. In block matrix multiplication, the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are partitioned to small blocks of size $s \times s$ where s is the dimension of small blocks usually decided based on the register size of the processor. We use $\boldsymbol{A}_{ij} \in \mathbb{R}^{s \times s}$ to denote the block at i^{th} row and j^{th} column in $\boldsymbol{A}$. Using the above data partitioning, the block matrix multiplication is shown in Algorithm 1.

Algorithm 1 Block Matrix Multiplication

```
Input: Input matrics $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$;
Output: Output matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$
for $i \leftarrow 1$ to $\frac{n}{s}$ do
    for $j \leftarrow 1$ to $\frac{n}{s}$ do
        Initialize $\boldsymbol{C}_{ij}$ in the registers
        for $k \leftarrow 1$ to $\frac{n}{s}$ do
            Load $\boldsymbol{A}_{ik}$ from on-chip memory
            Load $\boldsymbol{A}_{kj}$ from on-chip memory
            $\boldsymbol{C}_{ij} = \boldsymbol{C}_{ij} + \boldsymbol{A}_{ik} \times \boldsymbol{A}_{kj}$
        Store $\boldsymbol{C}_{ij}$ back to the external memory
```

B. Assumptions

To evaluate the energy efficiency of E-SRAM and O-SRAM, we propose the baseline E-SRAM system (Figure 2) and the O-SRAM wafer-scale system (Figure 3). The proposed system are organized based on the properties of E-SRAM and O-SRAM. We make the following assumptions:

- The input matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be fully stored in the on-chip memory of E-SRAM/O-SRAM based systems.
- In the O-SRAM based system, the on-chip optical memory resides in optical domain and the computation units (PE arrays) are in electrical domain. Two domains are connected through the optical-to-electrical interface.
- In the E-SRAM based system, the E-SRAM and PE array operate at the same clock frequency.
- In the O-SRAM based system, the optical on-chip memory has very high operating frequency and can send input data to multiple PE arrays concurrently. For example, the optical memory can operate at 20 GHz [24] while the PE array in electrical domain can operate at 500 MHz.

C. *Proposed system*

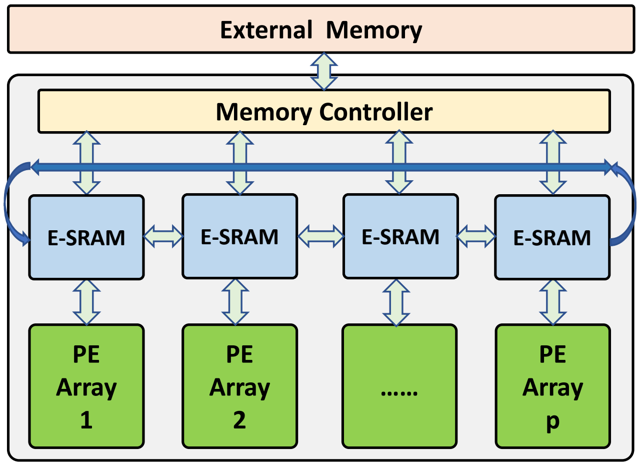

Fig. 2: The abstraction of E-SRAM based system.

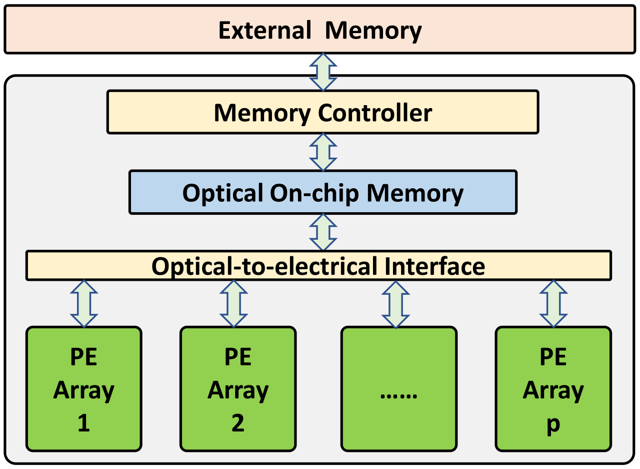

Fig. 3: The abstraction of O-SRAM based wafer-scale system.

E-SRAM based system: In the E-SRAM based system, the on-chip memory is logically arranged as 1-D ring structure. For example, on Cerebras WSE-2 wafer-scale system, the 1-D ring of on-chip memory can be easily formed through the on-wafer interconnection [28] of the processors. Each E-SRAM on the 1-D ring is connected to a Processing Element (PE) Array. The PE array is organized as 2-D systolic array, which will be elaborated later. The reasons for organizing the on-chip memory as the 1-D ring are two-fold: (1) E-SRAM has much lower operating frequency than O-SRAM. By distributing the E-SRAM into multiple blocks on the 1-D ring, the multiple E-SRAM blocks are able to achieve the same memory bandwidth as the single O-SRAM. (2) The PE arrays with 1-D ring on-chip memory can perform the block matrix multiplication efficiently using Algorithm 2.

Algorithm 2 Block Matrix Multiplication on 1-D ring

```
Input: Input matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$; $p$: number of PE arrays;
Output: Output matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$
Partition $\boldsymbol{A}$ and $\boldsymbol{B}$ into $p \times p$ partitions
Assign $\boldsymbol{A}_{i1}, \boldsymbol{A}_{i2}, ..., \boldsymbol{A}_{ip}$ to E-SRAM[$i$], $i = 1, 2, 3, ..., p$
Assign $\boldsymbol{B}_{1i}, \boldsymbol{B}_{2i}, ..., \boldsymbol{B}_{pi}$ to E-SRAM[$i$], $i = 1, 2, 3, ..., p$
for $i \leftarrow 1$ to $p$ do
    E-SRAM[$i$] sends its partitions of $\boldsymbol{B}$ to E-SRAM[$(i+1)\%N$]
    E-SRAM[$i$] receives the data from E-SRAM[$(i-1)\%N$]
    for $m \leftarrow 1$ to $p$ Parallel do                ▷ PE array $m$
        $j \leftarrow (m+i)\%p$
        for $k \leftarrow 1$ to $p$ do
            # execute $\boldsymbol{A}_{mk} \times \boldsymbol{B}_{kj}$ based on small block size $s \times s$
            $\boldsymbol{C}_{mj} = \boldsymbol{C}_{mj} + \boldsymbol{A}_{mk} \times \boldsymbol{B}_{kj}$
        Store $\boldsymbol{C}_{mj}$ back to the external memory
```

O-SRAM based wafer-scale system: In the O-SRAM based wafer-scale system, the on-chip memory is logically arranged as a one block of memory. Note, unlike E-SRAM, O-SRAM arrays communicate data through optical waveguides that are suitable for long distance communication at ultra-high speed making a single logical block of memory array feasible. In contrast, E-SRAM logical blocks are usually implemented in smaller sub-arrays to limit the length of metal wires and reduce parasitic capacitances and resistances to ensure high clock speed. The optical memory is connected to the PE arrays through the optical-to-electrical interface. Due to the extremely high frequency/memory bandwidth of the optical memory, it can serve the data requests from multiple parallel PE arrays concurrently. To execute the block matrix multiplication, the p parallel PE arrays can calculate p output blocks $\boldsymbol{C}_{ij}$ concurrently.

Algorithm 3 Block Matrix Multiplication on O-SRAM based wafer-scale system

```
Input: Input matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$; $p$: number of PE arrays;
Output: Output matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$
for $d \leftarrow 1$ to $\frac{n}{s\times p}$ do
    for $m \leftarrow 1$ to $p$ parallel do                ▷ PE array $m$
        $i = (d-1) \times p + m$
        for $j \leftarrow 1$ to $\frac{n}{s}$ do
            Initialize $\boldsymbol{C}_{ij}$ in the registers of PE array
            for $k \leftarrow 1$ to $\frac{n}{s}$ do
                Load $\boldsymbol{A}_{ik}$ from the optical memory
                Load $\boldsymbol{A}_{kj}$ from the optical memory
                $\boldsymbol{C}_{ij} = \boldsymbol{C}_{ij} + \boldsymbol{A}_{ik} \times \boldsymbol{B}_{kj}$
            Store $\boldsymbol{C}_{ij}$ back to the external memory
```

D. Hardware Modules

Processing Element (PE) Array: In both E-SRAM based and O-SRAM based system, the PE array is logically arranged as the 2-D systolic array (Figure 4), which is an efficient architecture for matrix-matrix multiplication. Since the 2-D systolic array has fully localized interconnection, it can be easily formed by the processors on the wafer-scale system (e.g., Cerebras WSE-2). Each PE is a Multiply-Accumulation (MAC) Unit. Suppose the PE array has the size of $p_{sys} \times p_{sys}$ and it uses the output stationary dataflow [5]. Therefore, the PE array can execute p_{sys}^2 multiplication-accumulation operation per clock cycle and each data will be reused for p_{sys} times in the systolic array.

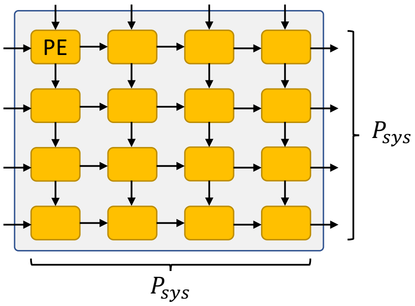

Fig. 4: The diagram of PE array

Organization of On-chip Memory: In the E-SRAM based system, each E-SRAM connected to a PE array has the data port of $2 \times p_{sys} \times d$ bits, where d is the bit width of a single data element. The E-SRAM can output $2 \times p_{sys}$ data elements per clock cycle to satisfy the data requirement of a PE array. In the O-SRAM based system, the single optical on-chip memory has data port width of $2 \times p_{sys} \times d$ bits. Since the optical memory has very high clock frequency, it can serve the data requests of multiple PE arrays concurrently. Suppose the optical on-chip memory in optical domain has the frequency of f_{op} and the PE array in electrical domain has the frequency of f_{el}, the data requests of $\frac{f_{\text{op}}}{f_{\text{el}}}$ PE arrays can be served by the optical on-chip memory concurrently.

Optical-to-electrical interface: Leveraging the differential readout of O-SRAM in [24], we designed an electro-optic sense amplifier at 22nm Globalfoundries node that can operate at a frequency of 20GHz. The power-performance metrics were scale to get an estimate for operation at 12nm node. The electro-optic sense amplifier was designed as a two stage amplifier, where the first stage acts as a pre-amplifier consisting of photodiodes that convert incoming light to differential electrical signals and the second stage is a high-speed dynamic comparator [29] that generates high or low electrical voltage in response to a input differential signal. Based on 12nm node, the energy consumption for sensing and conversion of a single bit of optical data to electrical data was estimated to be $5.26 \times 10^{-3} pJ$.

IV. Performance Model

In this section, we perform detailed theoretical analysis of the energy efficiency for the two systems. We use the following notations:

- Size of the input matrices $\boldsymbol{A}$,$\boldsymbol{B}$: $n \times n$; The bit width of each data element: d.
- The number of PE arrays in both systems: p; The dimension of each PE array: $p_{sys} \times p_{sys}$.
- The switching energy consumption of accessing 1-bit data from E-SRAM: $E_{\text{eletrical}}^{\text{switching}}$; The switching energy consumption of accessing 1-bit data from O-SRAM: $E_{\text{optical}}^{\text{switching}}$; The switching energy consumption of transferring 1-bit data from optical domain to electrical domain: E_{op2el}.
- The total energy consumption of p PE arrays to execute $\boldsymbol{A} \times \boldsymbol{B}$: $E_{\text{PE-arrays}}$

To execute the block matrix multiplication on PE arrays, the input matrices $\boldsymbol{A}$, $\boldsymbol{B}$ are partitioned to small blocks of size $p_{sys} \times p_{sys}$. Executing the block matrix multiplication of $\boldsymbol{A}$ and $\boldsymbol{B}$ involves $2n^3$ data accesses (n^3 data accesses to A and n_3 data accesses to B) if there is no data reuse. Since each data element will be reused for p_{sys} times in the PE array, there will be $\frac{2n^3}{p_{sys}}$ data accesses to the on-chip memory in total.

Energy consumption on E-SRAM based system: In E-SRAM based system, there are $\frac{2n^3}{p_{sys}}$ data accesses to the on-chip memory. Moreover, according to line 5-6 of Algorithm 2, there are $n^2 p$ data accesses due to the data communication among the p parallel E-SRAM blocks. Therefore, the overall energy consumption of E-SRAM based system is:

$$\begin{aligned} E_{\text{E-SRAM}} &= E_{\text{E-SRAM}}^{\text{static}} + E_{\text{E-SRAM}}^{\text{switching}} + E_{\text{PE-arrays}} \\ E_{\text{E-SRAM}}^{\text{static}} &= S_{\text{E-SRAM}} \times \frac{n^3}{p \times p_{sys}^2} \times \frac{1}{f_{pe}} \times P_{\text{electrical}}^{\text{static}} \\ E_{\text{E-SRAM}}^{\text{switching}} &= (\frac{2n^3}{p_{sys}} + n^2 p) \times d \times E_{\text{electrical}}^{\text{switching}} \end{aligned} \tag{1}$$

where $E_{\text{E-SRAM}}^{\text{static}}$ is the memory static energy consumption, $S_{\text{E-SRAM}}$ is the total size of E-SRAM, f_{pe} is the frequency of PE, $\frac{n^3}{p \times p_{sys}^2} \times \frac{1}{f_{pe}}$ denotes the total execution time and $P_{\text{electrical}}^{\text{static}}$ denotes the per-bit static power of E-SRAM. $P_{\text{electrical}}^{\text{static}}$ can be calculated by $P_{\text{electrical}}^{\text{static}} = E_{\text{electrical}}^{\text{static}} \times f_{pe}$ where $E_{\text{electrical}}^{\text{static}} \times f_{pe}$ is the per-bit static energy consumption of E-SRAM.

Energy consumption on O-SRAM based system: In O-SRAM based system, there are $\frac{2n^3}{p_{sys}}$ data accesses to the on-chip memory. Moreover, accessing the optical memory needs to go through the optical-to-eletrical interface, leading to additional energy consumption. Therefore, the overall energy consumption of O-SRAM based system is:

$$\begin{aligned} E_{\text{O-SRAM}} &= E_{\text{O-SRAM}}^{\text{static}} + E_{\text{O-SRAM}}^{\text{switching}} + E_{\text{PE-arrays}} \\ E_{\text{O-SRAM}}^{\text{static}} &= S_{\text{O-SRAM}} \times \frac{n^3}{p \times p_{sys}^2} \times \frac{1}{f_{pe}} \times P_{\text{optical}}^{\text{static}} \\ E_{\text{O-SRAM}}^{\text{switching}} &= \frac{2n^3}{p_{sys}} \times d \times (E_{\text{optical}}^{\text{switching}} + E_{\text{op2el}}) \end{aligned} \tag{2}$$

where $S_{\text{O-SRAM}}$ is the total size of O-SRAM and $P_{\text{optical}}^{\text{static}}$ denotes the per-bit static power of O-SRAM.

V. Evaluation

A. Experimental Setup

The power, performance estimates for O-SRAM was obtained from electro-optic simulations using *Lumerical Interconnect* [30]. Estimates for E-SRAM was based on SRAM design in Globalfoundries 12nm node, compute and PE array primitives were synthesized to obtain power-performance-area estimates at 12nm Globalfoundries PDK. Finally, SPICE simulations were used to obtain the energy estimate for the optical-to-electrical interface.

Using the above technology, the frequency of the electrical domain is $f_{\text{el}} = 500$ MHz and the frequency of the optical domain is $f_{\text{el}} = 20000$ MHz. We set the size of single PE array to be $p_{sys} \times p_{sys} = 16 \times 16$. For both two systems, we set number of PE arrays to be $p = \frac{f_{\text{op}}}{f_{\text{el}}} = 40$. The data width is set as $d = 32$.

TABLE I: The area of two systems

	On-chip Memory	PE Array	Total
E-SRAM system	$9.77\ mm^2$ (4 MB)	$5.92\ mm^2$	$15.69\ mm^2$
O-SRAM system	$3.84 \times 10^4\ mm^2$ (2 MB)	$5.92\ mm^2$	$3.84 \times 10^4\ mm^2$

Area: We assume that the wafer-scale O-SRAM system is implemented on the standard $300mm$ wafer, where the area is 70,650 mm^2. The area of 1-bit optical memory cell is $2400\mu m^2$. A single $300mm$ wafer can be deployed with up to $3.5MB$ optical memory. The area of 1-bit eletrical memory cell is $0.305\mu m^2$. A single O-SRAM bit-cell occupies the same area as 1KB of E-SRAM. Each PE array takes $1.48 \times 10^5 \mu m^2$ area. We assume the optical on-chip memory is 2 MB that stores input matrices $\boldsymbol{A}$,$\boldsymbol{B}$ of size 512×512. The total E-SRAMs in the E-SRAM based system is 4 MB for double buffering (used for data communication within 1-D ring structure). The area of two systems are shown in Table I. Note that the wafer-scale O-SRAM based system has much larger area than the E-SRAM based system. Since our objective is to compare the energy efficiency, and the two systems using same size pf PE arrays and similar amount of on-chip memory, the comparison is fair. Further, note that unlike E-SRAMs systems, O-SRAMs are amenable to large wafer-scale chips due to feasibility in long distance optical transfer of data. See Section VI about the area of optical memory in the future.

Energy Consumption: The energy consumption of two memory devices are shown in Table II and III. Table II and III demonstrate the per-bit energy consumption denoting the energy consumption for accessing a single bit of data. Compared with electrical memory device, the optical memory device has very small switching power. The PE array has operating voltage $V_{dd} = 0.3V$ and works under the temperature $25°C$. For each PE (MAC) in the PE array to operate on two 32-bit input data at 500 MHz frequency, the energy consumption is $1.1pJ$. In O-SRAM based system, the per-bit energy consumption of optical-to-electrical interface is $5.26 \times 10^{-3} pJ$.

TABLE II: Energy consumption of optical memory (Note that optical memory device has electrical part and optical part. Therefore, the static/switching energy consumption has electrical/optical part.)

Per-bit energy Consumption (pJ/bit)			
Static power		**Switching power**	
Electrical	**Optical**	**Electrical**	**Optical**
2.5×10^{-6}	1.67×10^{-6}	1.04	3.5×10^{-5}

TABLE III: Energy consumption of electrical memory (The static power consumption is calculated based on the frequency $f_{\text{el}} = 500$ MHz)

Per-bit Energy Consumption (pJ/bit)	
Static power	**Switching power**
1.175×10^{-6}	4.68

B. Comparison of Energy Efficiency

To compare the energy efficiency, the two systems execute the matrix multiplication $\boldsymbol{A} \times \boldsymbol{B}$ with matrix size of 512×512 ($n = 512$), which can be fully stored in the on-chip memory. The breakdown energy consumption are shown in Figure 5. The PE arrays on the two systems consume the same amount of energy (around $1.85 \times 10^7 pJ$). On the E-SRAM based system, the on-chip optical memory consumes $4.08 \times 10^9 pJ$ energy. On the O-SRAM based system, the on-chip electrical memory consumes $5.91 \times 10^8 pJ$ energy. The O-SRAM is $7.27\times$ more energy efficient than the E-SRAM. Considering the energy consumption of PE arrays, the O-SRAM based system is $7.07\times$ more energy efficient than the E-SRAM based system.

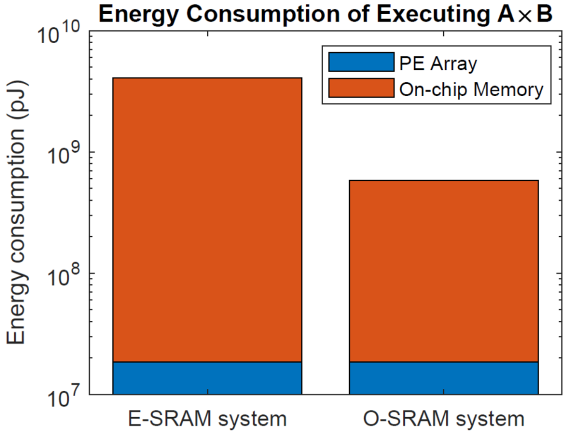

Fig. 5: The comparison of energy consumption (y-axis is in log-scale)

Break down energy consumption: Figure 6 and Figure 7 show the break down energy consumption of E-SRAM or O-SRAM, respectively. The E-SRAM has static energy consumption of 37.6 pJ and switching energy consumption of 4.08×10^9 pJ. Compared with the static energy consumption, the switching energy consumption is negligible for E-SRAM. The O-SRAM has static energy consumption of 66.7 pJ, switching energy consumption of 5.61×10^8 and 2.81×10^6 energy consumption of optical-to-eletrical memory interface.

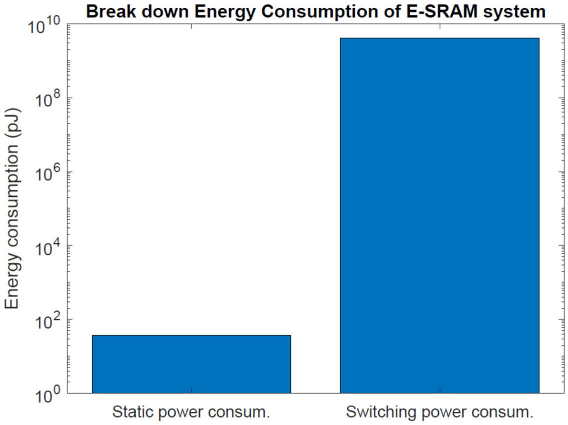

Fig. 6: Break down energy consumption of E-SRAM (y-axis is in log-scale)

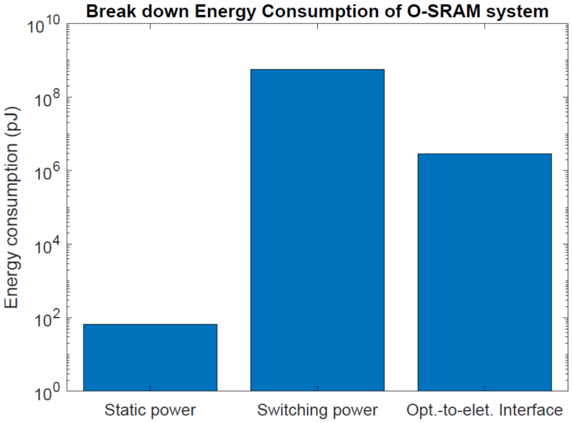

Fig. 7: Break down energy consumption of O-SRAM (y-axis is in log-scale)

VI. Discussion

Through the modeling of energy consumption, we demonstrate that optical memory has very low energy consumption on wafer-scale system for GEMM compared with electrical memory. The main reason for the low energy consumption is that the O-SRAM considered in this work uses reverse biased photodiodes and ring-resonator consuming minimal static power dissipation. Additionally, the ultra-high speed of O-SRAM also helps in reducing the energy consumption. In optical memory, the main energy consumption is the energy consumption of optical-to-electrical interface.

Remark on area of optical memory: Although optical memory has very low energy consumption, it occupies significant area compared with the electrical, limiting the density of the optical memory. Various approaches can be used to lower the area overhead of optical memory, including storing multi-bit data in a single bit-cell using techniques of wavelength division multiplexing and exploration of novel emerging optical memory materials and devices [31], [32] and ensuring their compatibility with commercial silicon foundry processes.

VII. Conclusion

Optical memory systems have been explored for a long time and is considered as a promising technology to achieve memory access speeds beyond state-of-the-art electrical memories. Recent advances in optical memory technologies have made it imperative to quantify system-level benefits of such ultra high-speed, energy-efficient, but area-expensive memories for complex compute operations. In this paper, we perform comprehensive modeling of the energy consumption of O-SRAM based system for GEMM. The evaluation results show that using the state-of-the-art optical memory technology, the O-SRAM based system is 7× more energy efficient than the E-SRAM based system.

Acknowledgment

This work is supported by the National Science Foundation (NSF) under grants OAC-2209563 and CNS-2009057.

References

[1] C. Yang, S. Chen, J. Zhang, Z. Lv, and Z. Wang, "A novel dsp architecture for scientific computing and deep learning," *IEEE Access*, vol. 7, pp. 36 413–36 425, 2019.

[2] T. Moreau, T. Chen, L. Vega, J. Roesch, E. Yan, L. Zheng, J. Fromm, Z. Jiang, L. Ceze, C. Guestrin *et al.*, "A hardware–software blueprint for flexible deep learning specialization," *IEEE Micro*, vol. 39, no. 5, pp. 8–16, 2019.

[3] B. Zhang, H. Zeng, and V. Prasanna, "Hardware acceleration of large scale gcn inference," in *2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 2020, pp. 61–68.

[4] B. Zhang, R. Kannan, and V. Prasanna, "Boostgcn: A framework for optimizing gcn inference on fpga," in *2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2021, pp. 29–39.

[5] Z.-G. Liu, P. N. Whatmough, and M. Mattina, "Systolic tensor array: An efficient structured-sparse gemm accelerator for mobile cnn inference," *IEEE Computer Architecture Letters*, vol. 19, no. 1, pp. 34–37, 2020.

[6] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[8] E. Wang, Q. Zhang, B. Shen, G. Zhang, X. Lu, Q. Wu, and Y. Wang, "Intel math kernel library," in *High-Performance Computing on the Intel® Xeon Phi™*. Springer, 2014, pp. 167–188.

[9] D. Kirk *et al.*, "Nvidia cuda software and gpu parallel computing architecture," in *ISMM*, vol. 7, 2007, pp. 103–104.

[10] "Amd xilinx gemx." [Online]. Available: https://github.com/Xilinx/gemx

[11] J. de Fine Licht, G. Kwasniewski, and T. Hoefler, "Flexible communication avoiding matrix multiplication on fpga with high-level synthesis," in *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2020, pp. 244–254.

[12] H. Giefers, R. Polig, and C. Hagleitner, “Measuring and modeling the power consumption of energy-efficient fpga coprocessors for gemm and fft,” *Journal of Signal Processing Systems*, vol. 85, no. 3, pp. 307–323, 2016.

[13] “Google environmental report 2021.” [Online]. Available: https://www.gstatic.com/gumdrop/sustainability/google-2021-environmental-report.pdf

[14] T. Alexoudi, G. T. Kanellos, and N. Pleros, “Optical ram and integrated optical memories: a survey,” *Light: Science & Applications*, vol. 9, no. 1, pp. 1–16, 2020.

[15] N. Pleros, D. Apostolopoulos, D. Petrantonakis, C. Stamatiadis, and H. Avramopoulos, “Optical static RAM cell,” *IEEE Photonics Technology Letters*, vol. 21, no. 2, pp. 73–75, 2008.

[16] A. Tsakyridis, T. Alexoudi, A. Miliou, N. Pleros, and C. Vagionas, “10 Gb/s optical random access memory (RAM) cell,” *Optics Letters*, vol. 44, no. 7, pp. 1821–1824, 2019.

[17] B. Dong, H. Cai, Y. Gu, Z. Yang, Y. Jin, Y. Hao, D. Kwong, and A. Liu, “Nano-optomechanical static random access memory (SRAM),” in *2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*. IEEE, 2015, pp. 49–52.

[18] B. Li, M. I. Memon, G. Mezosi, Z. Wang, M. Sorel, and S. Yu, “Optical static random access memory cell using an integrated semiconductor ring laser,” in *2009 International Conference on Photonics in Switching*. IEEE, 2009, pp. 1–2.

[19] T. Alexoudi, D. Fitsios, A. Bazin, P. Monnier, R. Raj, A. Miliou, G. T. Kanellos, N. Pleros, and F. Raineri, “III–V-on-Si photonic crystal nanocavity laser technology for optical static random access memories,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 22, no. 6, pp. 295–304, 2016.

[20] S. Pitris, C. Vagionas, T. Tekin, R. Broeke, G. Kanellos, and N. Pleros, “WDM-enabled optical RAM at 5 Gb/s using a monolithic inp flip-flop chip,” *IEEE Photonics Journal*, vol. 8, no. 2, pp. 1–7, 2016.

[21] Y. Liu, R. McDougall, M. Hill, G. Maxwell, S. Zhang, R. Harmon, F. Huijskens, L. Rivers, H. Dorren, and A. Poustie, “Packaged and hybrid integrated all-optical flip-flop memory,” *Electronics Letters*, vol. 42, no. 24, pp. 1399–1400, 2006.

[22] A. Trita, G. Mezosi, M. Zanola, M. Sorel, P. Ghelfi, A. Bogoni, and G. Giuliani, “Monolithic all-optical set-reset flip-flop operating at 10 Gb/s,” *IEEE Photonics Technology Letters*, vol. 25, no. 24, pp. 2408–2411, 2013.

[23] S. Wijeratne, A. Jaiswal, A. P. Jacob, B. Zhang, and V. Prasanna, “Performance modeling sparse mttkrp using optical static random access memory on fpga,” 2022. [Online]. Available: https://arxiv.org/abs/2208.10593

[24] R. Kudalippalliyalil, S. Chandran, A. P. Jacob, and A. Jaiswal, “Towards scalable, energy-efficient and ultra-fast optical sram,” *arXiv preprint arXiv:2111.13682*, 2021.

[25] G. Lauterbach, “The path to successful wafer-scale integration: The cerebras story,” *IEEE Micro*, vol. 41, no. 6, pp. 52–57, 2021.

[26] A. Brace, M. Salim, V. Subbiah, H. Ma, M. Emani, A. Trifa, A. R. Clyde, C. Adams, T. Uram, H. Yoo *et al.*, “Stream-ai-md: Streaming ai-driven adaptive molecular simulations for heterogeneous computing platforms,” in *Proceedings of the Platform for Advanced Scientific Computing Conference*, 2021, pp. 1–13.

[27] A. Kosson, V. Chiley, A. Venigalla, J. Hestness, and U. Koster, “Pipelined backpropagation at scale: training large models without batches,” *Proceedings of Machine Learning and Systems*, vol. 3, pp. 479–501, 2021.

[28] M. Jacquelin, M. Araya-Polo, and J. Meng, “Massively scalable stencil algorithm,” *arXiv preprint arXiv:2204.03775*, 2022.

[29] Y. Cao and C. Zhang, “Design of high speed dynamic comparator in 28nm cmos,” in *2018 IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia)*. IEEE, 2018, pp. 1–4.

[30] “Lumerical.” [Online]. Available: https://www.lumerical.com/products/interconnect/

[31] J. Geler-Kremer, F. Eltes, P. Stark, A. Sharma, D. Caimi, B. J. Offrein, J. Fompeyrine, and S. Abel, “A non-volatile optical memory in silicon photonics,” in *2021 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2021, pp. 1–3.

[32] J.-Y. Chen, L. He, J.-P. Wang, and M. Li, “All-optical switching of magnetic tunnel junctions with single subpicosecond laser pulses,” *Physical Review Applied*, vol. 7, no. 2, p. 021001, 2017.