Optical Comb-Based Monolithic Photonic-Electronic Accelerators for Self-Attention Computation

Tzu-Chien Hsueh, *Senior Member, IEEE*, Yeshaiahu Fainman, *Life Fellow, IEEE*, and Bill Lin, *Member, IEEE*

***Abstract*—This paper adopts advanced monolithic silicon-photonics integrated-circuits manufacturing capabilities to realize system-on-chip photonic-electronic linear-algebra accelerators for self-attention computation in various applications of deep-learning neural networks and Large Language Models. With the features of holistic co-design approaches, optical comb-based broadband modulations, and consecutive matrix-multiplication architecture, the system/circuit/device-level simulations of the proposed accelerator can achieve 2.14-TMAC/s/mm^2 computation density and 27.9-fJ/MAC energy efficiency with practical considerations of power/area overhead due to photonic-electronic on-chip conversions, integrations, and calibrations.**

***Index Terms*—Frequency comb, large language model, linear algebra, matrix-matrix multiplication, matrix-vector multiplication, micro-resonator, monolithic integration, racetrack resonator, self-attention, silicon photonics, transformer model.**

I. Introduction

RECENT advances in Large Language Models (LLMs) [1], and its underlying deep machine learning model called the Transformer [2], have shown unprecedented capabilities in artificial intelligence (AI). ChatGPT (or simply GPT), an important and impressive LLM developed by OpenAI [1], has the ability to understand and generate text that appears to resemble human conversations. According to a recent study [3], ChatGPT has already become the fastest-growing consumer application in history, having reached over 100 million active monthly users in just two months after launch. Beyond the most obvious impact of ChatGPT in its potential to transform communication and information, it is also capable of writing code in popular languages such as Python, C, Java, and JavaScript [4]. Also, the latest version of GPT was able to score near the 90-th percentile on an actual bar exam [5]. Beyond language-based applications, attention-head mechanism [2] has been applied successfully to scientific problems that on the surface seem quite far from the initial intent of Transformer models developed for natural language processing. One notable example application is the protein-folding problem that is well-known to require astronomical amounts of computing power using traditional computational chemistry approaches. In particular, protein-folding aims to predict precise three-dimensional structures of proteins, which is essential for understanding their function, designing new drugs, and advancing our understanding of diseases and biological systems. In a groundbreaking paper [6], the authors demonstrated AlphaFold, which is able to predict three-dimensional protein structures with remarkable accuracy as a generative AI task, based on the same underlying generative model that powers LLMs.

To realize Transformer models in the hardware physical layers with such high computation capacities as mentioned, photonic computing via monolithic silicon-photonics (SiPh) fabrication and integration [7], [8] is the primary implementation strategy of the proposed linear-algebra accelerator based on the major evidence as follows; in particular, this paper focuses on the acceleration of self-attention computation, which is the key bottleneck in Transformer models. First, since modern data bandwidth requirements and the standardization of SiPh integrated circuits (IC), photonic technology has been widely used for both long- and short-reach high-volume data communications [9], [10], [11], [12], [13], [14]. Meanwhile, due to the advancement of deep learning models far outpacing Moore's law [15] and energy/area bottleneck of classical von Neumann computing architectures, photonic devices and ICs, possessing inherent parallelism, high degree of connectivity, and speed-of-light propagation associated with wavelength-division-multiplexing (WDM) technique [16] in the optical communication, have been broadly adopted in the computation tasks of linear operations such as passive Fourier transforms [17] and matrix operations [15], [18], [19], [20], [21], [22], [23], [24], [25] exhibiting superior advantages of photonic computing in bandwidth density, processing latency, silicon area, and energy consumption. Especially, the concept of optical comb generations [26] plays a key role in the broadband incoherent photo-detection to cover more than 128 carriers with 30 GHz to 80 GHz frequency spacing across the entire WDM spectrum for the proposed linear-algebra accelerator. Second, the availability of commercial monolithic SiPh semiconductor process technology, e.g., GlobalFoundries 45SPCLO [7], [8], represents an exciting opportunity to explore holistic co-design approaches that leverage unique capabilities of photonics and CMOS electronics to advance the state of computing that is currently at a crossroad. Particularly, monolithic SiPh technology consolidating all required photonic and electronic devices/circuits for the proposed linear-algebra accelerator on a single die can tremendously eliminate

Manuscript received 22 November 2023; revised 30 March 2024 and 3 July 2024; accepted 3 July 2024. Date of publication 9 July 2024; date of current version 8 August 2024. This work was supported in part by National Science Foundation under Award 2023730, Award 2025752, Award 2045935, Award 2217453, and Award 2410053 and in part by ASML/Cymer Corporation. *(Corresponding author: Tzu-Chien Hsueh.)*

The authors are with the Department of Electrical and Computer Engineering, University of California, La Jolla, CA 92093 USA (e-mail: tzhsueh@ucsd.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2024.3425456.

Digital Object Identifier 10.1109/JSTQE.2024.3425456

power/area/integration overhead due to interfacing I/O circuits (SerDes and data transceivers), electrostatic-discharge (ESD) protection diodes, chip bumps/pads, and interposers/packages among separate photonic and electronic dies, which mostly have been ignored in the performance metrics of photonic computing literature [15], [19], [20], [21], [22], [23], [24], [25] but exposed by the limited computation scales and dimensions of their realistic hardware demonstrations.

The goal of the proposed monolithic photonic-electronic (MPE) linear-algebra accelerators is to exploit a new application of optical combs and practically realize a well-interfaced and power/area-efficient system-on-chip (SoC) possessing the functionality of high-dimensional matrix-vector multiplications (MVM), matrix-matrix multiplications (MMM), and double matrix-matrix multiplications (D-MMM) for the self-attention mechanism in Transformer models. The remainder of the paper is organized as follows. The background of the self-attention mechanism in Transformer models is summarized in Section II. The motivation for using monolithic SiPh technology and the self-attention optimization for the MPE realization are elaborated in Section III. The architecture and building-block functionality with simulated specifications of an MPE-MVM accelerator are analyzed in Section IV with practical considerations for on-chip comb-line power equalizations, nonlinearity/offset cancellations, process-voltage-temperature (PVT) variation calibrations, and circuits/devices noise tolerances. The architecture scalability and parallelism of MPE-MMM and MPE-D-MMM accelerators are described in Sections V and VI, respectively. The performance evaluation and conclusion are summarized in Section VII.

II. Background of Self-Attention

The Transformer model [2] is a type of deep learning network architecture that employs self-attention as a mechanism for processing input sequences. It can efficiently capture long-range dependencies and relationships within the data by differentially weighting the significance of each part of the input sequence. The Transformer model achieves this by means of a building block called a *scaled dot-product attention unit* that calculates attention weights simultaneously between all input tokens. This calculation produces embeddings for every token in context that contains information about the token itself, along with a weighted combination of other relevant tokens based on their respective attention weights. In particular, the Transformer model learns three weight matrices for each attention unit: the query weights W_Q, the key weights W_K, and the value weights W_V. The Query, Key, and Value matrices can then be generated by multiplying the input sequence $X = [x_0, x_1, \ldots, x_n]$ with the corresponding trained weight matrices:

$$Q = X \cdot W_Q; K = X \cdot W_K; V = X \cdot W_V \tag{1}$$

Given these matrices, the attention score can be calculated using the following scaled dot-product attention equation:

$$Attention\ (Q, K, V) = \ softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{2}$$

where d_k represents the dimension of the keys and values. As shown in (1) and (2), the central computations involve high-dimensional MVMs and MMMs.

The calculations in (1) and (2) basically form a single *attention-head*, with the corresponding set of weight matrices (W_Q, W_K, W_V). In the Transformer model proposed in [2], each layer has *multiple* attention-heads. With multiple attention-heads, each head can *attend* to a different notion of *relevance* by learning a different set of projection matrices (W_Q^i, W_K^i, W_V^i) for each attention-head i. The computations for each attention-head can then be executed in parallel to enable rapid processing. The outputs of h attention-heads are then concatenated and passed into a feedforward network layer as follows:

$$\begin{aligned} &Multihead\,(Q, K, V) \\ &= Concat\left[Attention\left(X \cdot W_Q^i, X \cdot W_K^i, X \cdot W_V^i\right)\right] \cdot W_O \end{aligned} \tag{3}$$

where W_O is the final projection matrix for the entire multi-headed attention unit. The overall Transformer architecture makes use of these multi-headed attention units as basic building blocks for encoders and decoders. Encoders and decoders can then be stacked together to provide increasing learning capacities to capture long-range dependencies and relationships within the input sequence.

III. Optimization Between Attention-Heads & Monolithic Photonic-Electronic Accelerators

Though a number of developments have explored photonic accelerators for convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [19], [20], [21], [22], [23], [24], [25] with promising results, many substantial challenges must be overcome when considering photonic computing for the attention-head calculations in Transformer workloads with such high-dimensional MMM computations. First, prior works on photonic accelerators for CNNs and RNNs have focused on MVM computations that are typically performed with *static* weights that do not require reprogramming once learned. However, the compute-kernels in a Transformer network are different: Transformer models require MMMs between the Query, Key, and Value matrices that are *dynamically* generated from *changing* inputs. Second, the attention-head calculations involve not only high-dimensional but also consecutive MMMs. If implemented in a naïve way with MVMs, the linear projection of inputs and the calculation of the scaled dot-product attention could require a significant number of conversions between the photonic and electronic domains. Third, prior works on photonic accelerators have largely assumed separate chip implementations for the photonic and electronic parts – these approaches incur very expensive chip-to-chip communications and integrations that decrease the effectiveness of a photonic computing approach to deep learning acceleration.

The innovations and approaches in Sections IV, V, and VI aim to address these substantial challenges by performing MVMs, MMMs, and D-MMMs purely in the photonic domain of a single monolithic SiPh chip without "intermediate" optical memory and O/E/O conversions to demonstrate the feasibility and capabilities of the MPE linear-algebra accelerator for

executing Transformer-model workloads with two layers of power/area/speed enhancements: first, the computation end-to-end E/O and O/E cost reduction through the monolithic SiPh integration; second, the number of O/E/O reduction through the attention-head architecture innovation, which is algorithmically elaborated in the rest of this section.

As shown in (1) and (2), the attention-heads involve a significant number of MMM computations, including the linear projections of the input vector by three trained matrices to produce Query, Key, and Value matrices, followed by a MMM of the Query and Key matrices and another MMM with the Value matrix. This paper proposed to co-design the attention-head computations in a holistic manner together with the capability of the MPE linear-algebra accelerator as mentioned. The first idea is to collapse the linear projections of the Query and Key matrices and their multiplication. In particular, $(Q \cdot K^T)$ can be rewritten as follows:

$$\begin{aligned} C = Q \cdot K^T &= (X \cdot W_Q) \cdot (X \cdot W_K)^T \\ &= \left[X \cdot \left(W_Q \cdot W_K^T\right)\right] \cdot X^T = (X \cdot W_C) \cdot X^T \end{aligned} \tag{4}$$

where $W_C = W_Q \cdot W_K^T$ can be computed offline since the entries in both W_Q and W_K are constants once trained. As shown in (4), collapsing the linear projections results in a *double matrix-matrix multiplication* (D-MMM), which can be realized by an MPE-D-MMM accelerator architecture described in Section VI, *without intermediate O/E/O conversions* between two individual MMMs. Once matrix C in (4) is computed as a complete D-MMM result, the outcome is converted from the photonic to the electronic domain, and then the scaling by $1/\sqrt{d_k}$ and $softmax(\cdot)$ required in each attention-head can be efficiently realized as simple digital shifts and table lookup operations in the electronic domain. In particular, $\sqrt{d_k}$ is typically chosen to be some 2^m so scaling $c_{ij} \in C$ by $1/\sqrt{d_k}$ can be performed simply by a right-shift of c_{ij} by m bits. As shown in [27] and [28], the $softmax(\cdot)$ function can be rewritten as follows:

$$\begin{aligned} softmax\left(a_i\right) &= \frac{e^{a_i}}{\sum_{j=1}^{K} e^{a_j}} \\ &= \exp\left[a_i - a_{max} - \log\left(\sum_{j=1}^{K} e^{a_j - a_{max}}\right)\right] \end{aligned} \tag{5}$$

where a_{max} is the maximum among the a_i values. Also, as discussed in [27] and [28], $\exp(\cdot)$ and $\log(\cdot)$ can be efficiently approximated with table lookups, where the quantization of the a_i values is co-designed with the MPE accelerator to optimize for the best tradeoffs. With matrix $S = softmax(C/\sqrt{d_k})$ computed as right-shifts and table lookups in accordance with (5), the remaining attention-head computation is another D-MMM of three matrices:

$$Attention\left(Q, K, V\right) = S \cdot V = S \cdot (X \cdot W_V) \tag{6}$$

which can again take advantage of the MPE-D-MMM accelerator to avoid the intermediate O/E/O conversion within the operation of (6).

IV. MPE Matrix-Vector Multiplications

A unified MPE-MVM accelerator serves as the key building block of the MPE-MMM, MPE-D-MMM, and eventually attention-heads in the Transformer model with high degrees of reconfigurability in terms of the internal matrix weights and accelerator-to-accelerator on-chip physical connections. The primary MVM functionality is realized by exploiting high degrees of freedom in spatial parallelism with the WDM technique [16]. As shown in Fig. 1, each high-dimensional MPE-MVM accelerator consists of "d" B-bit high-speed digital-to-analog converters (HS-DAC), "d" comb-line power equalization DACs (EQ-DAC), "d^2" low-power static DACs (LP-DAC), "d" transimpedance amplifiers (TIA) individually followed by "d" B-bit analog-to-digital converters (ADC), digital registers/logics, clock distribution, and discrete-time iteration mechanism in the electronic domain as well as "d" vector micro-ring modulators (z-MRM) for the input vector E/O conversion, "d" comb-line power equalization micro-ring modulators (e-MRM) for rectifying all input light-wave powers, "d^2" matrix micro-ring modulators (y-MRM) for the matrix E/O conversion with photonic WDM-based MVM operations, "d" germanium (Ge) photodetectors (PD), optical power splitter (O-PS), and waveguides in the photonic domain.

The input d-by-1 data vector is denoted by $\mathrm{Z_{d\times 1}}$ with elements $\mathrm{z_i}$, $\mathrm{i} = 1 \sim \mathrm{d}$; the output d-by-1 data vector is denoted by $\mathrm{YZ_{d\times 1}}$ with elements $\mathrm{yz_i}$, $\mathrm{i} = 1 \sim \mathrm{d}$ (note that each "yz" presents a "single-word" variable, NOT "y times z"); the primary d-by-d matrix is denoted by $\mathrm{Y_{d\times d}}$ with elements $\mathrm{y_{ij}}$, both i and $\mathrm{j} = 1 \sim \mathrm{d}$ individually. The MVM functionality is represented by $\mathrm{Y_{d\times d} \cdot Z_{d\times 1} = YZ_{d\times 1}}$. The timings of both input and output vectors are managed by the electronic circuits in the digital domain to enable the flexibility of cascading and parallelizing the MPE-MVM accelerators and, therefore, to perform multi-stage MVMs while the MPE-MVM accelerator can also seamlessly interface with other on-chip digital circuits, processors, lookup tables, registers, and memory. The detailed circuits and interconnection within the MPE-MVM accelerator are shown in Fig. 2 as a zoom-in version of the first MVM row from the input $\mathrm{z_1}$ to output $\mathrm{yz_1}$. It is important to note that the clock-rate per MVM operation, playing one of the key roles for the computation throughput (MAC/s) [15], is determined by the circuits/devices latency from the clock edge launching the HS-DAC through the MRMs, O-PS, WDM-based MVM operation, PD, TIA, all the way to the ADC output data sampled by the next clock edge, which sets the clock period per MVM computation. Thus, the speeds of all electronic and photonic circuits/devices within this critical path determine the clock period (i.e., register-to-register clock cycle) and eventually computation throughput (MAC/s) of this MPE-MVM accelerator, especially the electronic circuits dominating the entire power/area/speed performance metrics.

A. WDM-Based MVM Computation Architecture

The MVM functionality is established on the concept of WDM incoherent data transmission in optical communication systems [9], [12], [16]. The data carriers are "d" of laser light-waves with individual wavelengths $\mathrm{\lambda_i}$, $\mathrm{i} = 1 \sim \mathrm{d}$, and

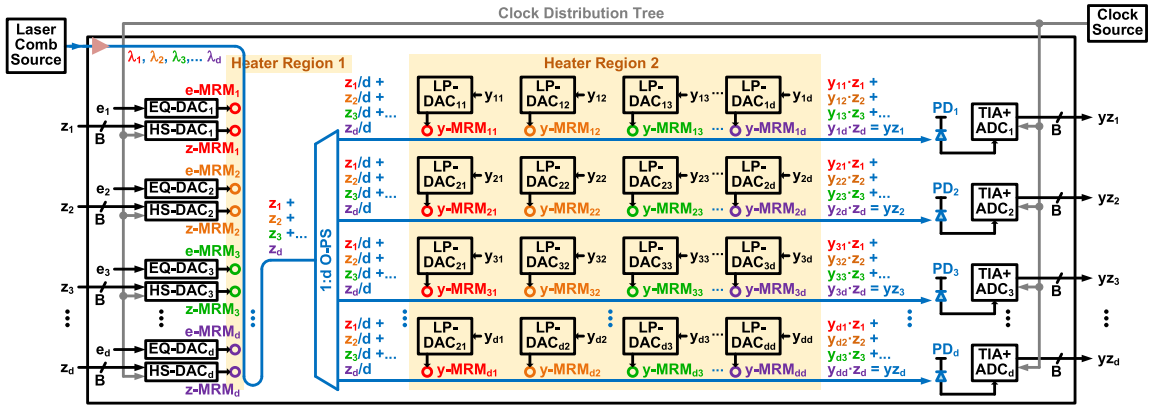

Fig. 1. The system block diagram of the MPE-MVM linear-algebra accelerator. Note that each "yz" presents a "single-word" variable, NOT "y times z".

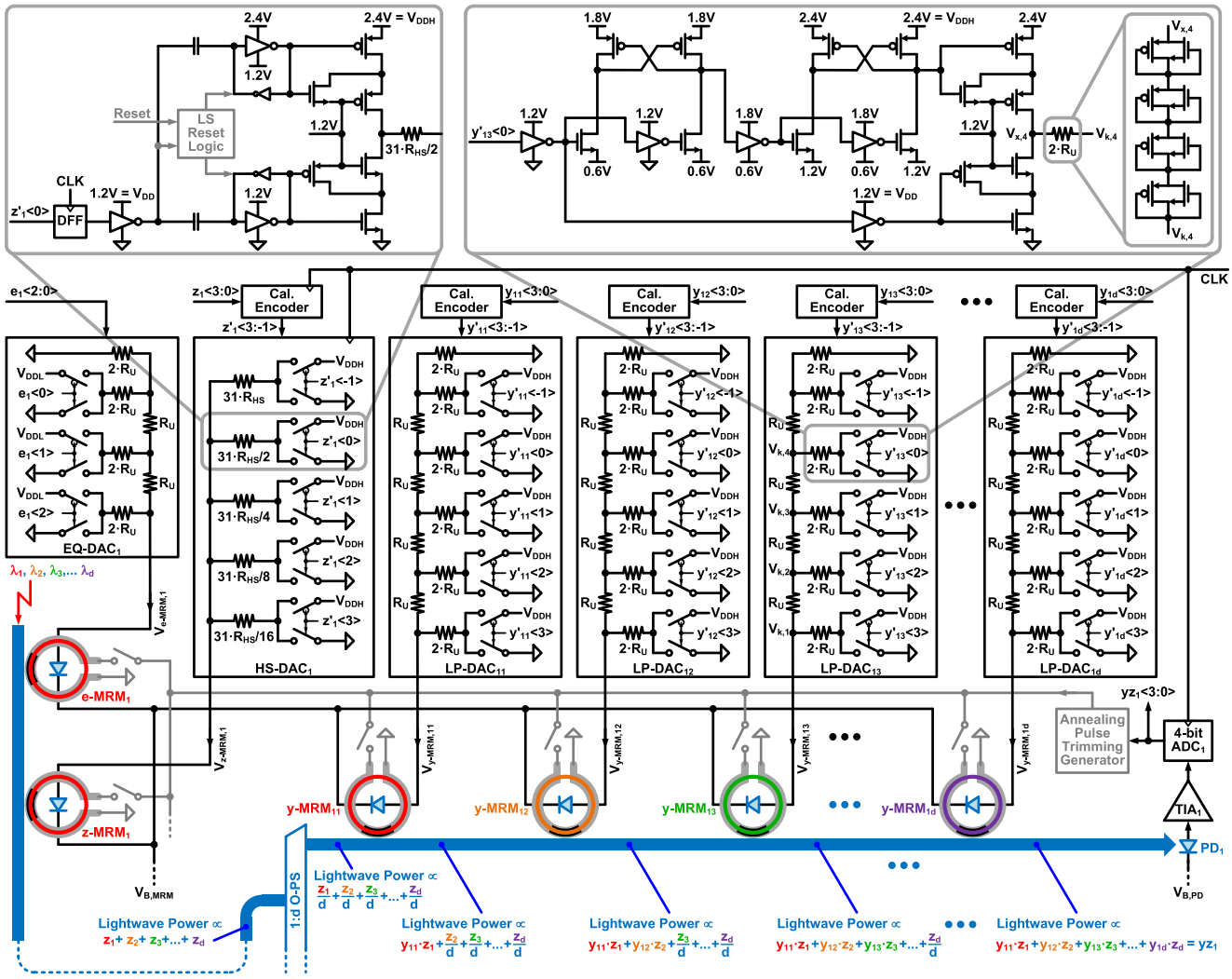

Fig. 2. The zoom-in version of the first MPE-MVM row from the input z_1<3:0> to output yz_1<3:0> with detailed circuits, interconnections, and clocking relation among the HS-DAC, EQ-DAC, LP-DACs, MRMs, PD, TIA, ADC, and post-fabrication trimming mechanism.

proper wavelength spacing $\Delta\lambda_{WDM,i}$, $i = 1 \sim d$, injected from an external (off-chip) laser comb source through an on-chip grating coupler into a single waveguide as shown in the left of Fig. 1, where each wavelength λ_i is modulated twice on the input waveguide by e-MRM$_i$ and z-MRM$_i$ for comb-line power equalization and E/O conversion, respectively. Each e-MRM$_i$ is driven by the comb-line power equalization code e_i<2:0> through its EQ-DAC$_i$, which is a low dynamic-range (DR) 3-bit R2R-DAC elaborated in Section IV-C, to slightly attenuate the light-wave power of λ_i so that all wavelengths can reach almost identical power calibrated through ADC$_i$ after the thermal tuning process described in Section IV-G. has been completed.

After the comb-line power equalization for all wavelengths λ_i, $i = 1 \sim d$, each digital z_i data element of the input vector $Z_{d\times1}$ is D/A and E/O converted through HS-DAC$_i$ and z-MRM$_i$, respectively, so that z_i data can be carried and presented by the λ_i light-wave power in the photonic domain. More precisely, the λ_i light-wave power becomes linearly proportional to z_i with a consistent scalar across all $i = 1 \sim d$ as shown in Fig. 1 and bottom-left of Fig. 2, where the aggregate light-wave power is proportional to $\sum_{i=1}^{d} z_i = (z_1 + z_2 + z_3 + \cdots + z_d)$. Then, the O-PS evenly splits this aggregate light-wave power into its "d" fan-outs so that the light-wave powers in the waveguides of all MVM rows shall be identical and proportional to $\sum_{j=1}^{d} z_j/d = (z_1 + z_2 + z_3 + \cdots + z_d)/d$. Note that the wavelength index for the i-th MVM row is "j" not "i".

After the O-PS, the light-wave power contributed by all λ_j, $j = 1 \sim d$, in the i-th MVM row will sequentially go through the resonance effects of "d" MRMs (i.e., y-MRM$_{ij}$, $j = 1 \sim d$). Similar to the previous scenario, each y-MRM$_{ij}$ controlled by y_{ij} can only affect or modulate the λ_j light-wave power ($\propto z_j/d$) when the indexes j of y_{ij} and z_j are matched. As shown in the bottom of Fig. 2 for the first row of the MVM operation ($i = 1$), for example, all λ_j light-wave powers ($\propto \sum_{j=1}^{d} z_j/d$) together pass through the resonance effect of y-MRM$_{11}$ ($j = 1$), but only λ_1 light-wave power ($\propto z_1/d$) gets modulated to be proportional to $(y_{11}\cdot z_1)$; the rest of λ_j, $j = 2 \sim d$, light-wave powers can be all-pass filtered through y-MRM$_{11}$ without any power change. By resonating through all y-MRM$_{1j}$, $j = 1 \sim d$, all λ_j light-wave powers in the first MVM row can be respectively modulated according to y_{1j}, $j = 1 \sim d$, at the speed of light. Right before reaching PD$_1$, the total power on the first waveguide becomes proportional to $\sum_{j=1}^{d} y_{1j}\cdot z_j = (y_{11}\cdot z_1 + y_{12}\cdot z_2 + y_{13}\cdot z_3 + \cdots + y_{1d}\cdot z_d) = yz_1$, which is equivalent to the dot-product presentation of the input vector $Z_{d\times1}$ and the first-row vector of matrix $Y_{d\times d}$. By replicating the same process across all row vectors of $Y_{d\times d}$ in parallel, the total light-wave power of each waveguide can be eventually proportional to yz_i with a consistent scalar across all $i = 1 \sim d$. After the following O/E and A/D conversions through PD$_i$ and ADC$_i$, respectively, the MVM operation is completed, and all elements (yz_i, $i = 1 \sim d$) of the output vector $YZ_{d\times1}$ are preserved in the electronic domain.

The similarity between WDM communication and WDM computation is that both rely on multiple wavelengths to independently carry their own signal/data information and to simultaneously propagate through a common communication channel to maximize the communication capacity or computation parallelism. On the other hand, there are two major differences. First, each light-wave power is only modulated once in WDM communication, which is merely for E/O conversion, but in WDM computation, each light-wave power requires to be modulated at least "twice" to perform the E/O conversions and equivalent effect of "multiplications." Second, the receiver side of WDM communication needs to distinguish and separate all wavelengths and then detect each light-wave power individually to recover the data carried by each wavelength. Though WDM computation doesn't need to explicitly separate all wavelengths, the wavelength spectrum spacing still needs to be maintained because the "multiplication" is done when all light-wave powers are simultaneously present in the waveguide. Also, the equivalent summation of the dot-product requires consistent power-absorption photo-detections across the entire WDM spectrum to maintain the computation integrity (or accuracy) during the O/E conversions. These two major differences between WDM communication vs. WDM computation dominantly determine the achievable E/O/E modulation/detection speeds ($\geq$16 Gb/s vs. $\leq$4 Gb/s) and D/A/D data resolutions ($\leq$2 bits vs. $\geq$4 bits) of these two optical WDM-based systems.

B. High-Speed Digital-to-Analog Data Converters

In the primary computation path, each digital z_i data element of the input data vector $Z_{d\times1}$ represented by B-bit digital data (e.g., 4-bit in Fig. 2) is firstly converted into an analog-voltage level through HS-DAC$_i$ to drive a z-MRM$_i$ for modulating the light-wave power of λ_i as the E/O conversion process. To accommodate a reasonable E/O DR, the non-linear transmission power-gain induced by the MRM across all possible 2^B voltage levels (e.g., $2^4 = 16$ levels in Fig. 2) can be alleviated by adding one calibration bit in the HS-DAC$_i$, so a basic digital calibration encoder is required for HS-DAC$_i$ mapping the incoming B-bit z_i data to (B+1)-bit z_i' data to perform better E/O conversion linearity while maintaining the same amount of 2^B voltage-levels (not 2^{B+1}), which is further discussed in Section IV-D.

Each HS-DAC is implemented by a large-swing voltage-mode driver architecture [29] to maintain up to GHz sampling-rates (i.e., clock-rates), low static-power consumption, and high-voltage driving capability for maximizing the light-wave power DR per MRM. As shown in Fig. 2, the binary segments controlled by z_1'<3:−1> in HS-DAC$_1$ have identical architecture, but the transistor-size (or driving capability) of each segment is reciprocally scaled according to its own series-resistor. In other words, all binary HS-DAC segments can be simply implemented by parallelizing multiple least-significant-bit (LSB) segments, as shown in the top-left corner of Fig. 2, in a manner of powers of 2. For example, the segment driven by z_1'<3> is formed by sixteen LSB segments driven by z_1'<−1>. Based on a certain binary combination of z_1'<3:−1>, some series resistors can be shorted to V_{DDH}, and the rest to GND; thus, the HS-DAC output can generate a voltage level based on the voltage divider formed by all the parallel pull-up and pushed-down resistors between V_{DDH} and GND. Note that two AC-coupled level-shifters are required in each HS-DAC for push-pull data latches

[9] to enable high-speed, low-power, low-latency single-stage 2× voltage level-shifting from 1.2-V regular digital supply to 2.4-V high-voltage supply. Overall, to simultaneously maximize the E/O DR up to 2.4 V and maintain 45-nm CMOS reliability in 45SPCLO, separating power domains of the push-pull latches and adding cascade transistors in the push-pull driver paths are necessary to confine all CMOS devices operating within (1.2 to 2.4) V or (0 to 1.2) V.

The speed of the HS-DAC is determined by two factors. First, the latency from the clock edge of the z_1'<3:−1> register (i.e., DFF) to the push-pull switching transistors of the voltage-mode drivers, containing the delays of DFF clock-to-Q, digital data buffers, level-shifting AC-coupling capacitors, and push-pull data latches, is about 5× of a 20-ps fan-out of 4 (FO4) logic delay in 45SPCLO, i.e., 100 ps. Second, the RC time-constant of the HS-DAC output is determined by all parallel series-resistors from the standpoint of AC signal, i.e., a roughly constant and data-independent resistance $\approx (31{\cdot}R_{HS}/16 \parallel 31{\cdot}R_{HS}/8 \parallel 31{\cdot}R_{HS}/4 \parallel 31{\cdot}R_{HS}/2 \parallel 31{\cdot}R_{HS}) = R_{HS}$, and all parasitic capacitance contributed by the transistors, resistors, and MRM, i.e., a roughly constant and data-independent capacitance $C_{HS\text{-}DAC} + C_{MRM} \approx 30$ fF [13]. For a 2-GSym/s, 4-bit, 2.4-V DR data, the (R_{HS}·30-fF) time-constant should be at least less than 58 ps so that the HS-DAC output can spend less than 200 ps settling to its static analog-voltage level within an LSB/2 of the 4-bit resolution, i.e., $\exp(-200\text{-ps}/58\text{-ps}) < 1/(2^{4+1})$. Therefore, R_{HS} needs to be less than 58-ps/30-fF ≈ 2 kΩ, which sets the static power consumption per HS-DAC discussed in the next paragraph. Overall, the E/O latency spends 300 ps (= 100-ps FO4 logic delays + 200-ps HS-DAC output settling time) out of the 500-ps clock-period budget (i.e., 2-GHz clock-rate).

The power consumption of each HS-DAC is contributed by both dynamic power and static power. First, the FO4 digital logics in each HS-DAC, including the DFFs, calibration encoder, data buffers, and push-pull data latches, mainly consume 0.2-mW dynamic power at 2-GSamp/s, which follows $C_L \cdot V_{DD}^2 \cdot f_{CLK}/2$ or $C_L \cdot (V_{DDH} - V_{DD})^2 \cdot f_{CLK}/2$ with the simulation of 50% Logic-1 and 50% Logic-0 data pattern per logic gate across multiple MVM operation cycles. Second, the static power consumption is mainly due to the DC current path from V_{DDH} to GND of the resistor-divider formed by the parallel series-resistors and voltage-mode drivers as mentioned. Though the AC resistance of this HS-DAC architecture is constant, the DC path resistance or current between the V_{DDH} and GND is data-dependent. By assuming the data patterns of z_i with corresponding voltage-levels are uniformly distributed between 0 to $(2^B - 1)$ across multiple MVM operation cycles, the average static power consumption per HS-DAC can be expressed as follows:

$$P_{HS\text{-}DAC,ST} = \frac{V_{DDH}^2}{R_{HS}} \cdot \frac{\sum_{k=1}^{2^{B-1}} (k-1) \cdot \left(2^B - k\right)}{2^{B-1} \cdot \left(2^B - 1\right)^2} \quad (7)$$

In the case of Fig. 2, where $B = 4$, $R_{HS} \approx 2$ kΩ, $V_{DDH} = 2.4$ V, each HS-DAC consumes 0.45-mW static power. Overall, the average power consumption of each HS-DAC is 0.65 mW, including both dynamic and static power.

The silicon area of each HS-DAC shown in Fig. 2 is around 100-μm×20-μm, including the voltage-mode driver, push-pull data latches, unsilicided poly-resistors for the voltage-divider, calibration encoder logics, and a static logic for resetting the data-dependent initial conditions of the push-pull data latches [9]. The 20-μm height per HS-DAC tile is designed to match the height per z-MRM tile.

C. Low-Power R2R Digital-to-Analog Data Converters

The approach to realizing multiplication of each vector and matrix elements $y_{ij}{\cdot}z_j$ (e.g., $y_{11}{\cdot}z_1$ in Fig. 2) is to provide a transmission power-gain $\propto y_{ij}$ on the top of the λ_j light-wave power pre-modulated to $\propto z_j/d$ in the i-th waveguide. That is, after the E/O conversion and O-PS for z_i becoming z_j/d in the i-th MVM row, the element-to-element "multiplication" is basically done by another light-wave power modulation through the y-MRM_{ij} only effective to the λ_j light-wave power. Thus, LP-DAC_{ij} is required to convert a digital multiplicand y_{ij} to its corresponding voltage-level for setting the power-gain of y-MRM_{ij}, which is the same D/A and E/O operations in Section IV-B. However, the DAC speed requirement for converting y_{ij} is not critical at all because the value or state of the matrix $Y_{d\times d}$ is pre-determined and static during the regular MVM operations as described in Section III for the attention-head computations. This fact beneficially allows to use low-speed but low-power small-area R2R voltage-divider architecture [30] to implement "d^2" of the LP-DACs in a high-dimensional MVM accelerator as shown in Fig. 1. Similarly, the EQ-DACs used for the incoming λ_i, $i = 1 \sim d$, power equalization can be implemented by the R2R architecture as well with a much smaller DR requirement due to the purpose of static power-gain controls.

To accommodate a reasonable E/O DR for y_{ij}, the non-linear transmission power-gain induced by y-MRM_{ij} across all possible 2^B voltage-levels (e.g., $2^4 = 16$ levels in Fig. 2) is alleviated by adding one calibration bit, and thus a digital calibration encoder is required for LP-DAC_{ij} mapping the incoming B-bit y_{ij} data to (B+1)-bit y_{ij}' data to perform better E/O conversion linearity while maintaining the same amount of 2^B voltage-levels (not 2^{B+1}). LP-DAC_{ij} shown in Fig. 2 is constructed by alternating series-R and parallel-2R resistances with push-pull switches to form a voltage-divider driving the capacitive electrode of y-MRM_{ij} based on the value of y_{ij}. According to the number of binary bits of y_{ij}, an LP-DAC can minimize the amount of the resistor-and-transistor components for generating all required voltage levels, which is very beneficial to both power and area savings by taking advantage of static operations or low-speed requirements.

The concept of each parallel-2R segment in the LP-DAC is similar to the bit-segment in an HS-DAC; both require a voltage-mode driver architecture using push-pull switching transistors to short $2{\cdot}R_U$ resistor to V_{DDH} or GND. However, because of the low-speed LP-DAC operation, the pull-up transistor can be driven by a two-stage static level-shifter as shown in the top-right of Fig. 2 to simultaneously maximize the E/O DR up to 2.4 V and maintain 45-nm CMOS reliability in 45SPCLO with negligible power consumption since all logics are under

static states during the MVM operations. The only noteworthy power consumption is static power in the voltage divider formed by the R2R network conducting a static current from V_{DDH} to GND. Before showing the expression of average static power consumption per LP-DAC, multiple indexes and variables need to be defined: "i" and "j" are still the row and column indexes of the matrix; by treating all LP-DAC$_{ij}$ are identical, "i" and "j" do not involve in the power calculation actually; $k = 1 \sim 2^B$ represents the voltage-level index; $p = 1 \sim B$ represents the circuit-node index, $q = 1 \sim B$ represents the Kirchhoff's Voltage Law (KVL) superposition index of each circuit-node; R_U is the unit-resistor shown in Fig. 2; $R_p = (G_p/H_p)\cdot R_U$ is the one-side equivalent resistance of each circuit-node; G_p and H_p are integers; G_p/H_p is the simplest fraction; $V_{k,p}$ is the KVL superposition voltage of each voltage-level and each circuit-node. By assuming the data patterns of y_{ij} and corresponding voltage-levels are uniformly distributed between 0 to $(2^B - 1)$ across the entire matrix $Y_{d\times d}$, the average static power consumption per LP-DAC can be expressed as follows:

$$
\begin{aligned}
&G_1 = 1,\ H_1 = 0 \Rightarrow R_1 = \infty, \quad p = 1\\
&R_p = R_{p-1} \,||\, (2R_U) + R_U = \frac{G_p}{H_p}\cdot R_U, \quad p = 2 \sim B\\
&\frac{V_{k,p}}{V_{DDH}} = \sum_{q=1}^{B} y_{ij,k}\,\langle B-q\rangle \cdot \frac{\min\left[G_p, G_q\right]}{2^{p+q-1}}, \quad \begin{matrix} k = 1 \sim 2^B \\ p = 1 \sim B \end{matrix}\\
&P_{R2R\text{-}DAC,ST} = \frac{V_{DDH}^2}{R_U}\cdot\\
&\frac{\sum_{k=1}^{2^B}\sum_{p=1}^{B} y_{ij,k}\,\langle B-p\rangle \cdot \left(1 - \frac{V_{k,p}}{V_{DDH}}\right)}{2^{B+1}}
\end{aligned}
\tag{8}
$$

In the case of Fig. 2, where $B = 4$, $R_U \approx 5\ M\Omega$, $V_{DDH} = 2.4$ V, each LP-DAC consumes about 7.2 μW. To reach up to about 5 MΩ with a small amount of silicon area for each R_U, the resistance template is implemented by multi-stacked sub-threshold-region transistors [31] as the example shown in the top-right of Fig. 2, where $2\cdot R_U$ is formed by four stacked P/N parallelized diode-connected transistors. Since the relative resistance ratios in the R2R network are more critical rather than the absolute resistance values, the process-corner variations and temperature coefficients of the active resistors are tolerable with proper transistor sizing under the 4-bit accuracy requirements. However, the nonlinearity due to data-dependent terminal voltages across the active resistors requires extra attention. For example, the resistance $2\cdot R_U$ between $V_{x,4}$ and $V_{k,4}$ in Fig. 2 is data-dependent since $V_{x,4}$ can be either V_{DDH} or GND, and $V_{k,4}$ can vary from GND to (85/128)$\cdot V_{DDH}$ based on the $V_{k,p}$ formula in (8). This paper proposed complementary sub-threshold-region P/N active-resistor architecture: when $V_{x,4} = V_{DDH} > V_{k,4}$, the sub-threshold bias conditions are $0 < |V_{GS,PMOS}| < V_{th,PMOS}$ and $V_{GS,NMOS} = 0$; when $V_{x,4} = GND < V_{k,4}$, the sub-threshold bias conditions are $0 < |V_{GS,NMOS}| < V_{th,NMOS}$ and $V_{GS,PMOS} = 0$. That is, under these two possible $V_{x,4}$ conditions, the PMOS and NMOS sub-threshold biases are automatically set to be complimentary, which can cancel the first-order resistance nonlinearity due to the data-dependent $V_{x,4}$. On the other hand, the resistance nonlinearity due to 2^B different $V_{k,p}$ values based on y_{ij}<3:0> in general can be calibrated by the additional bit in y_{ij}'<3:−1> together with the y-MRM$_{ij}$ nonlinear power-gain calibration discussed in Section IV-D.

Though the transistor numbers and resistance values in the LP-DAC all seem higher than those in the HS-DAC, the silicon area of each LP-DAC shown in Fig. 2 is around 20-μm × 20-μm, including the voltage-mode driver, push-pull data latches, sub-threshold-region active resistors, and calibration encoder logics. The 20-μm width per LP-DAC tile is designed to match the width per y-MRM tile.

D. Micro-Ring Modulators & E/O Conversions

Once a data element (z_i or y_{ij}) of the vector or matrix is converted and settled to an analog-voltage level through either an HS-DAC or LP-DAC at the P-type electrode of its z-MRM or y-MRM, the light-wave power of a certain wavelength, which is located within the MRM resonance bandwidth, can be effectively modulated according to the voltage delta between the N-type (i.e., the DC voltage from $V_{B,MRM}$) and P-type (i.e., the settled DC voltage from the DAC output) electrodes of the MRM (i.e., the reverse bias between the MRM P/N junction).

The resonance wavelength, coupling strength, radiative loss, and quality factor of an MRM are mainly determined by its transmission waveguide widths, ring waveguide width, transmission-to-ring waveguide gap, and especially ring radius, which is the dominant factor in setting the footprint of the MRM. The radius r_{ij} design of y-MRM$_{ij}$ (or r_i of z-MRM$_i$) in the whole WDM spectrum shall at least satisfy two criteria. First, the minimum free spectral range (FSR) is confined by all WDM isolation spacing $\Delta\lambda_{WDM,ij}$ between adjacent λ_{ij} and the number of λ_{ij} for the WDM (i.e., "d") [32] as shown in the top of Fig. 3(a). Second, each y-MRM$_{ij}$ resonance wavelength λ_{ij} under a particular resonance mode m_{ij} (an integer) corresponds to its effective refractive index $n_{eff}(\lambda_{ij})$, silicon propagation constant $\beta(\lambda_{ij})$, and ring circumference $L_{ij} = 2\pi\cdot r_{ij}$; in other words, when the resonance condition is satisfied in the MRM cavity (i.e., L_{ij}), a constructive interference is established by a certain wavelength having its round-trip phase shift equal to an integer multiple of 2π [33]. These two criteria are summarized in (9) and (10), respectively, as follows:

$$
\begin{aligned}
&\Delta\lambda_{FSR,ij} = \frac{\lambda_{ij}^2}{n_g(\lambda_{ij})\cdot 2\pi\cdot r_{ij}} \geq \sum_{j=1}^{d}\Delta\lambda_{WDM,ij}\\
&\Rightarrow r_{ij} \leq \frac{\lambda_{ij}^2}{n_g(\lambda_{ij})\cdot 2\pi\cdot\sum_{j=1}^{d}\Delta\lambda_{WDM,ij}}
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&2\pi\cdot m_{ij} = \beta(\lambda_{ij})\cdot L_{ij} = \frac{2\pi}{\lambda_{ij}}\cdot n_{eff}(\lambda_{ij})\cdot 2\pi\cdot r_{ij}\\
&\Rightarrow r_{ij} = \frac{\lambda_{ij}}{2\pi\cdot n_{eff}(\lambda_{ij})}\cdot m_{ij}
\end{aligned}
\tag{10}
$$

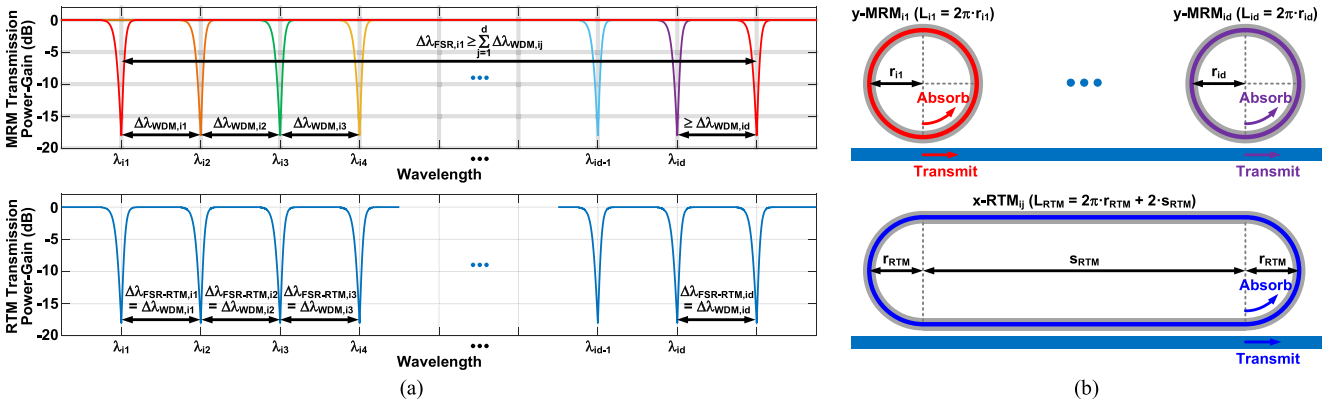

Fig. 3. (a) The transmission responses of y-MRM$_{ij}$, j = 1 ∼ d (or z-MRM$_i$, i = 1 ∼ d, if λ_{ij} is replaced with λ_i), and x-RTM$_{ij}$, i = 1 ∼ n, j = 1 ∼ d. (b) The conceptual geometry of y-MRM$_{ij}$ and x-RTM$_{ij}$. Note that the transmission response and geometry of x-RTM$_{ij}$ are not functions of i and j.

TABLE I
DESIGN TRADEOFF EXAMPLES AMONG MVM DIMENSIONS, WDM CROSSTALK, AND FABRICATION ERRORS

d	$n_{eff}(\lambda_{ij})$ j = 1 ∼ d	$n_g(\lambda_{ij})$ j = 1 ∼ d	$m_{RTM,ii}$ j = 1 ∼ d	L_{RTM}	λ_{ij} j = 1 ∼ d	$\Delta\lambda_{WDM,ii}$ j = 1 ∼ d	m_{ii} j = 1 ∼ d	r_{ii} j = 1 ∼ d	$\Delta r_{rms}/r_{ij}$ Error Scale
32	3.74 ∼ 3.73	5.02 ∼ 4.98	2321 ∼ 2290	951.32 μm	1534.5 ∼ 1550 nm	0.49 ∼ 0.51 nm	71 or 72	4.63 ∼ 4.76 μm	1×
32	3.76 ∼ 3.73	5.06 ∼ 4.98	1160 ∼ 1129	469.01 μm	1519.0 ∼ 1550 nm	0.97 ∼ 1.03 nm	35 or 36	2.25 ∼ 2.38 μm	2×
16	3.74 ∼ 3.73	5.02 ∼ 4.98	1160 ∼ 1145	475.66 μm	1535.0 ∼ 1550 nm	0.99 ∼ 1.01 nm	71 or 72	4.63 ∼ 4.76 μm	1×

where $\Delta\lambda_{FSR,ij}$ is the FSR of y-MRM$_{ij}$; $n_g(\lambda_{ij})$ is the silicon group index; $2\pi/\lambda_{ij}$ is the free-space propagation constant; λ_{ij} presents the free-space resonance wavelength of y-MRM$_{ij}$ although the light-waves propagate in the silicon. To describe the y-MRM$_{ij}$ (or z-MRM$_i$) design procedure, a practical example with realistic parameter values for determining the radius r_{ij} of y-MRM$_{ij}$ is demonstrated in the following paragraphs as the initial design without considering any post-process trimming or thermal control.

Typically, the nominal wavelength-spectrum bandwidth in the silicon of 45SPCLO is in the range of 1180 nm to 1550 nm [12]. Therefore, with the MRM quality-factor (i.e., Q) up to 10000 [12] in 45SPCLO, the WDM isolation spacing $\Delta\lambda_{WDM,ij}$ across 32 (= d) wavelengths is initially set to ∼0.5 nm, and the maximum WDM wavelength λ_{i32} is set to 1550 nm in this example. At this point, it is important to note that the exact 32 resonance wavelengths (λ_{ij}, j = 1 ∼ 32) distributed from 1534.5 nm to 1550 nm and 32 isolation spacing ($\Delta\lambda_{WDM,ij}$, j = 1 ∼ 32) slightly varying from 0.49 nm to 0.51 nm listed in Table I are actually determined by the design equation of the optical comb-based racetrack modulator (RTM) discussed in Section VI.

With all designated λ_{ij}, $\Delta\lambda_{WDM,ij}$, and their corresponding $n_g(\lambda_{ij})$, the FSRs of all y-MRM$_{ij}$ have to meet the criterion of $\Delta\lambda_{FSR,ij} \geq 32{\cdot}0.5$ nm = 16 nm, and then the upper bound of each r_{ij} can be specified based on (9). Finally, by properly choosing the mode integer m_{ij} with corresponding $n_{eff}(\lambda_{ij})$, all r_{ij} of y-MRM$_{ij}$ can be obtained and distributed from 4.63 μm to 4.76 μm according to (10). By following the same procedure and setting different values of "d" and initial $\Delta\lambda_{WDM,ij}$, the design tradeoffs among the MVM dimension, WDM crosstalk, and MRM fabrication error are summarized in Table I, including three different scenarios: the case one has the worst WDM crosstalk due to the smallest $\Delta\lambda_{WDM,ij}$; the case two has the worst MRM fabrication error $\propto (2\pi\cdot\Delta r_{rms})/(2\pi\cdot r_{ij})$ where $2\pi\cdot\Delta r_{rms}$ is the circumference error of y-MRM$_{ij}$ due to random process variation and independent of r_{ij}; the case three has the worst computation throughput because of the smallest values of "d" out of these three scenarios. In any case, the silicon area per MRM can be confined within a 20-μm × 20-μm tile, including the central ring area and peripheral keep-out halo.

The linearity of the entire analog circuit, photonic dynamic range, and signal conversions in the MPE accelerator is extremely crucial; this issue could continuously dominate the development of photonic computing. Specifically in the E/O conversions, the deterministic nonlinearity is mainly caused by the sigmoid-like high-Q power transmission response shown in Fig. 4(a) as the zoom-in version of Fig. 3(a) for y-MRM$_{11}$, simulated by a Verilog-A approach [34] matched with the MRM model in the 45SPCLO process design kit (PDK), even though the transmission resonance can be almost linearly shifted with the reverse bias driven by a linear DAC.

For the example in Fig. 4(a), the wavelength of λ_1-laser carrying data information in its light-wave power is designated at 1534.5 nm to match the resonance wavelength of y-MRM$_{11}$ with the reverse bias corresponding to y_{11}<3:0> = b'0000 = 0. Under 16 different reverse biases generated by a 4-bit linear DAC according to y_{11} varying from b'0000 = 0 to b'1111 = 15, the MRM resonance wavelength and transmission response are shifted horizontally at a constant rate of 0.04-nm/V. However, this causes the power attenuation (i.e., power gain <1) of λ_1-laser nonlinearly distributed across the vertical E/O DR as highlighted by black triangular marks in Fig. 4(a) and (b) on dB scale. The same y_{11} vs. E/O conversion curves are also

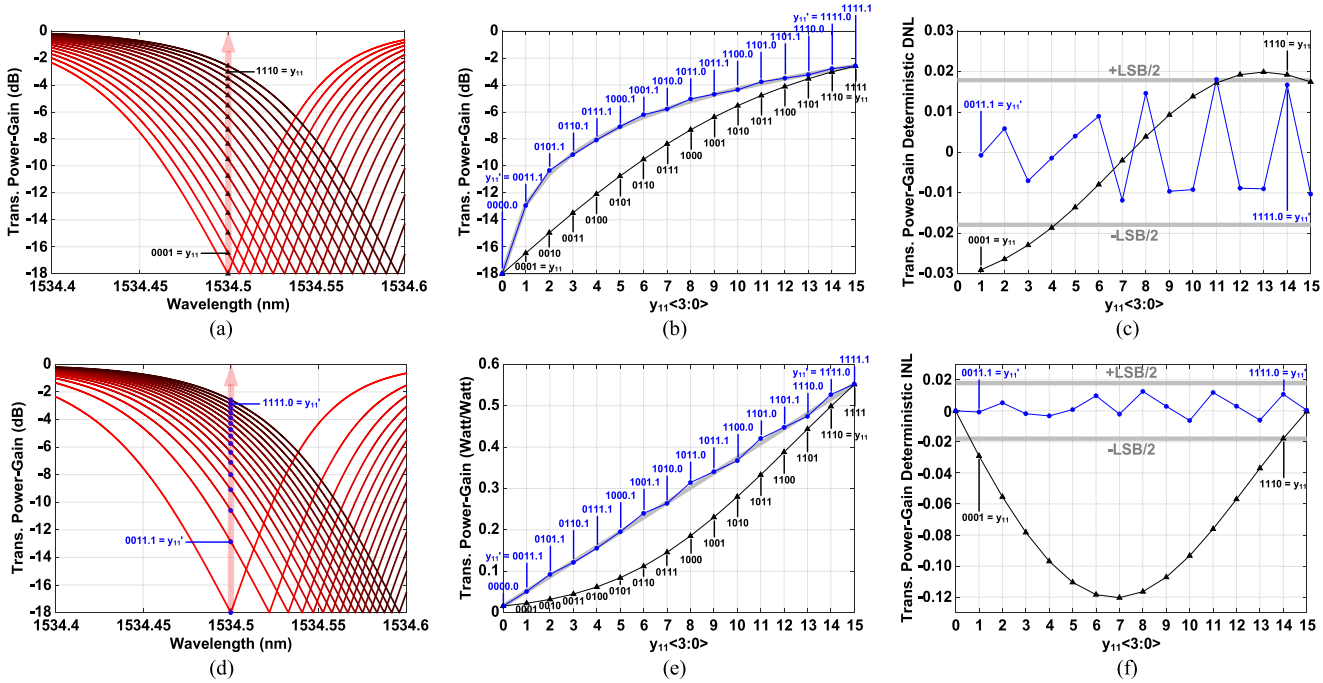

Fig. 4. (a) The transmission responses of y-MRM$_{11}$ modulated by a 4-bit linear DAC. (b) The transfer curves of y_{11}<3:0> vs. 4-bit nonlinear (black) and linearized (blue) E/O conversions on a dB scale. (c) The deterministic DNL of the 4-bit nonlinear (black) and linearized (blue) E/O conversions. (d) The transmission responses of y-MRM$_{11}$ modulated by a 4-bit nonlinear DAC for the linearization. (e) The identical transfer curves shown in (b) but on a linear scale. (f) The deterministic INL of the 4-bit nonlinear (black) and linearized (blue) E/O conversions.

shown in Fig. 4(e) but on a linear scale for the sake of linearity demonstration. To improve the linearity of the black curve, a 5-bit linear DAC with a 5-bit input y_{11}'<3:−1> is exploited to generate 32 different power gains within the same E/O DR. Since the MRM characteristics stay the same, either using a 4-bit DAC or a 5-bit DAC, the corresponding 16 or 32 power-gains are still located on top of the black curve in Fig. 4(e). However, the key of this linearization technique is that the system-level specification only requires 16 power gains, so the 5-bit DAC creates the freedom to properly select 16 (blue markers) out of 32 MRM power gains to fit the ideal linear line in gray. This selection process can be done by a digital encoder to map y_{11}<3:0> to y_{11}'<3:−1> after the black curve has been measured. For the example in Fig. 4(e), when y_{11}<3:0> = b'0100 = 4, the MRM power gain is about 0.06, which is far away from the ideal value of 0.15. Meanwhile, when y_{11}'<3:−1> = b'0111.1 = 7.5, the MRM power gain is indeed very close to the ideal value of 0.15. Therefore, the encoder can intentionally map y_{11}<3:0> = 4 to y_{11}'<3:−1> = 7.5 to cause the MRM power gain to be a value very close to 0.15. This encoding process turns out to select 16 of 32 y_{11}'<3:−1> values as well as corresponding 16 of 32 MRM power gains to form the blue curve in Fig. 4(e). Therefore, although the transmission response shifts nonlinearly according to y_{11}<3:0> due to the mapping process in Fig. 4(d), the power attenuation of λ_1-laser relatively has better linearity, which is very essential to the linear-algebra accelerator. The linearity enhancement of this technique can be qualified by the deterministic E/O differential nonlinearity (DNL) and integral nonlinearity (INL) in Fig. 4(c) and (f), respectively, where both DNL and INL are reduced down to the range within ±LSB/2 of the 4-bit E/O conversion accuracy when this technique is enabled. Note that DNL and INL [35] are commonly used in representing the linearity performance of signal converters, including DACs, ADCs, E/O, O/E, etc. In the case of Fig. 4(e), although it is easy to tell the blue curve has better linearity since it aligns with the ideal linear line in gray better than the black curve does, the DNL and INL values extracted from Fig. 4(e) and shown in Fig. 4(c) and (f), respectively, can specifically quantify the deviations of the blue and black curves from the ideal linear line in Fig. 4(e); i.e., the lower absolute values of DNL and INL represent lower deviations from the ideal case, lower nonlinearity, and, in other words, better linearity.

E. Transimpedance Amplifiers, Single-Ended-to-Differential Amplifiers & Analog-to-Digital Data Converters

Once each dot-product yz_i converted from the aggregate light-wave power to a photocurrent in the electronic domain through the Ge PD$_i$ with higher than 0.5-A/W responsivity and 35-GHz bandwidth in 45SPCLO [8], the following CMOS circuit blocks, including the TIA, single-end-to-differential amplifier (S2D-AMP), and ADC as shown in Fig. 5, further convert yz_i from the analog photocurrent to a B-bit digital word. Up to this point, the entire signal flow of the MVM operation is basically completed in terms of the end-to-end digital data format, computation outcome, and operation clock-cycle.

The TIA in Fig. 5 is implemented by a voltage-to-current feedback-amplifier architecture. The feedforward path is formed by a complimentary P/N transconductance ($G_{m,TIA} = g_{m,p} + g_{m,n}$) stage while both share the single DC bias current from

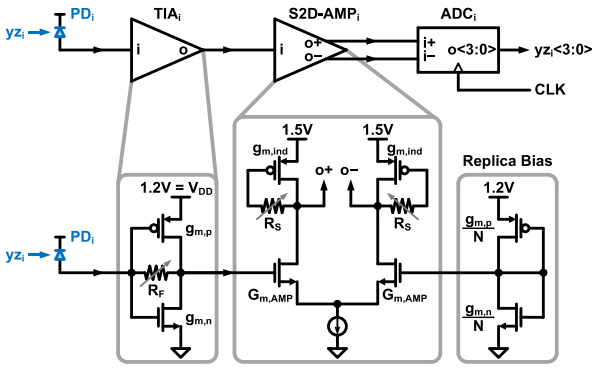

Fig. 5. The block diagram of each MPE-MVM dot-product O/E conversion circuit, and the schematics of TIA_i and $S2D\text{-}AMP_i$.

the TIA supply (V_{DD}) to GND. The feedback path is formed by a tunable resistor R_F, which plays a key role in setting the TIA gain, bandwidth, and input/output impedance across all different PVT conditions with $G_{m,TIA}$. In addition, this TIA is self-biased because of a zero DC feedback current through R_F and series P/N diode-connection; the input and output DC bias voltages can be self-locked at the operating point of $(V_{GS,n} + V_{SG,p}) = V_{DD}$ without requiring other biasing mechanism and extra power consumption. This TIA architecture has been broadly used in high-speed optical receiver front-ends [11], [36] because of its simplicity for both high-bandwidth and low-power characteristics. The simplified output resistance, transfer function, and average input-referred noise power spectrum density (PSD) [36] of this TIA architecture are shown in (11), (12), and (13), respectively:

$$R_{TIA} \approx \frac{R_F}{1 + G_{m,TIA} \cdot R_F} \tag{11}$$

$$TF_{TIA}(s) \approx -\frac{G_{m,TIA} \cdot R_F \cdot R_{TIA}}{1 + s \cdot R_{TIA} \cdot C_{TIA}} \tag{12}$$

$$\overline{I^2_{n,in,TIA}} \approx \frac{2\pi \cdot \kappa \cdot T}{R_F^2} \cdot \left(\frac{\gamma}{G_{m,TIA}} + \frac{1}{G^2_{m,TIA} \cdot R_{TIA}} \right) \tag{13}$$

where $\kappa = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant; T = 300 K is the thermal dynamic temperature in Kelvin; $\gamma \approx 2.5$ is the excess noise coefficient for deep submicron technology; $C_{TIA} \approx 30$ fF is the capacitive load at the TIA_i output, which is mainly contributed by the input capacitance of $S2D\text{-}AMP_i$. Based on the design example so far with the worst case 0.5-A/W PD_i responsivity and 670-μW DR of the aggregate WDM light-wave power, the DR of the TIA input photocurrent is 335 μA. To achieve 335-mV DR at the TIA output, $G_{m,TIA}$ and R_F are chosen to be 1 mA/V and 1.65 kΩ, respectively, so that the TIA DC gain can be about 335-mV/335-μA = 1 k$\Omega \approx |TF_{TIA}(0)|$ based on (12). Meanwhile, the TIA bandwidth $\approx 1/(2\pi \cdot R_{TIA} \cdot C_{TIA})$ reaches 8.52 GHz, which is sufficiently higher than the 1-GHz Nyquist frequency of the 2-GSym/s per yz_i data. More importantly, this design choice leads to the average input-referred noise PSD of this TIA as the first active-stage of the O/E interface lower than 6.26 pA/$\sqrt{\text{Hz}}$ based on (13). About the static power consumption, the TIA for its 1-mV/A $G_{m,TIA}$ and scaled replica generating identical DC common-mode voltages for the $S2D\text{-}AMP_i$ differential-pair together consume about 0.1 mW from the 1.2-V supply.

The second-stage, $S2D\text{-}AMP_i$, is implemented by a fundamental common-source differential amplifier with a pair of active-inductor loads to mainly convert and buffer the single-ended TIA output to a differential signal for minimizing asymmetrical kickbacks and common-mode noise at the input of the following ADC_i as shown in Fig. 5. The active-inductor load [37] is formed by a diode-connected PMOS with a tunable feedback resistor R_S to enhance the high-frequency gain for sharpening data symbol transitions and to intentionally unbalance the gains of the positive and negative outputs for cancelling the nonideal single-ended-to-differential conversion process due to the finite output impedance of the tail current source and PVT variations. Based on the same design example so far, if the $S2D\text{-}AMP_i$ output capacitive load C_{AMP} mainly contributed by the input capacitance of ADC_i is about 80 fF, its output resistance R_{AMP} is chosen to be 600 Ω so that the default output 3-dB bandwidth $\approx 1/(2\pi \cdot R_{AMP} \cdot C_{AMP})$ is about 3.32 GHz, which dominates speed of the entire MVM operation. Fortunately, without extra power consumption, the active-inductor load can effectively enhance the bandwidth up to 5.5 GHz, which is sufficient for the 1-GHz Nyquist frequency of the 2-GSym/s per 4-bit yz_i data. This is a good example to emphasize that, in the monolithic integrated MVM, the $S2D\text{-}AMP_i$ output (or ADC_i input) is usually the critical bandwidth node (e.g., 5.5 GHz) of the entire WDM signal path rather than the TIA output bandwidth (e.g., 8.52 GHz) in general. On the other hand, in other heterogeneously integrated MVMs, the critical bandwidth node could be at the TIA input, which is the interface between the separate electronic and photonic ICs requiring dedicated I/O bumps, ESDs, and passive-inductor peaking (T-coil) circuits. The DC voltage-gain $|TF_{AMP}(0)| = G_{m,AMP} \cdot R_{AMP}$ of $S2D\text{-}AMP_i$ is set to 3×, so that the 335-mV TIA output DR can be converted to 1-V_{diff} ADC input DR through $S2D\text{-}AMP_i$. Also, $G_{m,AMP}$ and the static power consumption of $S2D\text{-}AMP_i$ can be respectively determined as 5 mA/V and 0.75 mW from the 1.5-V supply.

The third-stage, ADC_i, is implemented by a B-bit flash ADC architecture consisting of $(2^B - 1)$ strong-arm-latch (SAL) based clocked-comparators [38] as shown in Fig. 6. The flash ADC architecture is suitable for high-speed and low-resolution applications with the downside of 2^B exponential-growth of the circuit area and dynamic power. In the case of this on-chip 4-bit O/E interface, the 15 SAL-based comparators per flash ADC actually offer an adequate compromise between area and conversion-rate (i.e., sampling-rate = 2 GS/s) without consuming static power. Each SAL-based clocked comparator consists of a SAL followed by an RS-latch to form an edge-sampled DFF as a full clock-cycle register. The dual differential pairs of each SAL are exploited to compare the voltage difference between its differential-input and differential-reference voltages. The SAL architecture allows itself to complete the signal integration, regeneration, and decision within a half-period of the sampling clock by using a single phase of the sampling

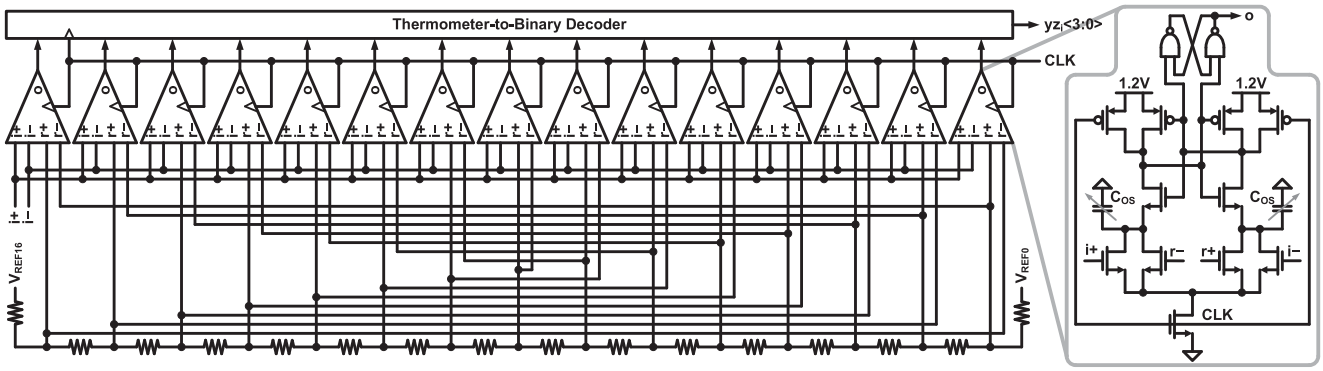

Fig. 6. The schematic of ADC_i in each MPE-MVM dot-product O/E conversion circuit, and the schematic of the SAL-based clocked comparator.

clock and the concept of integrating amplifications so that the average power consumption of 4-bit ADC is about 1.2 mW ($= 15 \times 80$-μW). Note that additional 3-bit capacitor-banks (C_{OS}) for SAL offset/nonlinearity cancellations across PVT variations and a common resistor-ladder for all SALs differential reference-voltage generations both consume negligible power.

The limitation of the low-power O/E circuit designs mentioned above is bounded by the A/D accuracy criterion along with the specification of the laser injection power for the WDM-based MVM operation. Therefore, the design iteration is necessary to consider the noise power from the primary contributors in this O/E interface, including the PD_i shot noise ($\overline{I^2_{n,PD}}$), TIA_i circuit output noise ($\overline{V^2_{n,TIA}}$), and $S2D\text{-}AMP_i$ circuit output noise ($\overline{V^2_{n,AMP}}$). The overall noise power at the ADC_i input can be expressed as follows:

$$\begin{aligned} P_{n,O/E} \approx & \int_0^\infty \overline{I^2_{n,PD}}(f) \cdot |TF_{TIA}(f)|^2 \cdot |TF_{AMP}(f)|^2 \cdot df \\ & + \int_0^\infty \overline{V^2_{n,TIA}}(f) \cdot |TF_{AMP}(f)|^2 \cdot df \\ & + \int_0^\infty \overline{V^2_{n,AMP}}(f) \cdot df \\ \approx & \, 0.5qI_{PD}\left(G^2_{m,TIA}R^2_F R^2_{TIA}\right) \\ & \times G^2_{m,AMP}R_{AMP}/C_{AMP} \\ & + kT\left(\gamma G_{m,TIA}R^2_{TIA} + R_{TIA}\right) \\ & \times G^2_{m,AMP}R_{AMP}/C_{AMP} \\ & + 2kT\Big(\gamma G_{m,AMP}R_{AMP} \\ & + \gamma g_{m,ind}R_{AMP} + 1\Big)/C_{AMP} \end{aligned} \tag{14}$$

where $q = 1.602 \times 10^{-19}$ C is the elementary charge; $g_{m,ind}$ is the transconductance of the PMOS used for the active-inductor load; $R_{AMP} \approx R_S/(1+g_{m,ind}\cdot R_S) \approx 600\ \Omega$ is the output resistance of $S2D\text{-}AMP_i$. In (14), the frequency-domain integrals are simplified by only considering the bandwidth of $S2D\text{-}AMP_i$ due to its bandwidth domination. Based on the circuit design parameters so far with the maximum $I_{PD} \approx 335$ μA, the overall noise power at the ADC_i input is around 11 μW. Compared to the quantization-noise power of the 4-bit flash ADC $= V^2_{LSB}/12 = (1\text{-}V_{diff}/15)^2/12 = 370$ μW, there is still a margin of 2.5 bits ≈ 15.3 dB $= 10\cdot\log_{10}(370\text{-}\mu\text{W}/11\text{-}\mu\text{W})$ to accommodate dark noise, flicker noise, supply noise, clock jitter, resistor-ladder noise, comparator noise, and residual offset, which are not included in the noise estimation of (14). The overall active area of TIA_i, $S2D\text{-}AMP_i$, and ADC_i is about 100-μm $\times$ 20-μm for the O/E conversion of each MVM row.

F. Optical Power Splitters & Laser Injection Power

The MPE-MVM accelerator as shown Fig. 1 requires external laser comb sources as being developed in [26] to interface with the optical grating coupling on the monolithic SiPh chip. After the on-chip comb-line power equalization and E/O conversion as described in Section IV-A, the O-PS duplicates the pre-summed vector elements into "d" identical waveguides ready for the following WDM-based MVM operation. To realize this high fan-out O-PS, this paper exploited the concept of adiabatic Y-junction power splitters possessing low-loss, high-bandwidth, high-polarization insensitivity, and high-tolerance to fabrication errors by using a nonlinear taper coupling technique to shorten the horizontal dimension of each 50/50 Y-junction power splitter [39]. In the case of a fan-out of "d" WDM-based MVM, the O-PS contains $\log_2(d)$ stages horizontally, and each stage is vertically formed by multiple 50/50 Y-junction power splitters in parallel of a number from 2^0 ($= 1$) to $2^{\log2(d)-1}$ ($= d/2$) for the first to last stages, respectively. The simulation result of the 50/50 Y-junction power slitter in Fig. 7(a) on the 160-nm silicon-on-insulator (SOI) 45SPCLO platform shows that the 17.5-μm nonlinear taper coupler within a total 35-μm footprint, including the length from the single horizontal fan-in to two horizontal fan-outs, can reach insertion losses less than 0.05 dB as shown in Fig. 7(b) consistent with the measurement data reported in [40] for the transverse electric (TE) mode in the range of 1530-nm to 1550-nm WDM spectrum as specified in Table I. Meanwhile, the negligible insertion-loss variations can help to minimize the DR overhead of EQ-DACs for the comb-line power equalization. To match the 20-μm height per MVM row, the overall silicon area of a 1-to-d O-PS is $[\log_2(d)\cdot 35\text{-}\mu\text{m}] \times [d\cdot 20\text{-}\mu\text{m}]$, which is 175-$\mu$m $\times$ 640-μm when $d = 32$.

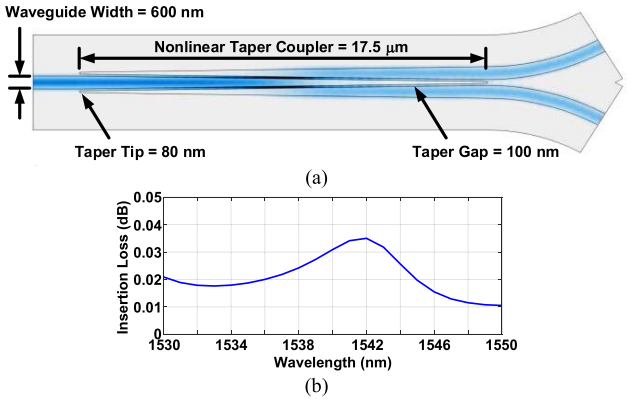

Fig. 7. (a) The mask layout, dimensions and (b) insertion loss of the 50/50 Y-junction power-splitter with the nonlinear taper coupling technique.

The aggregate signal power loss "$\log_2(d)\cdot(3.01 + 0.05)$-dB" of a 1-to-d O-PS is consolidated with the transmission losses of e-MRMs, z-MRMs, and y-MRMs in each MVM row to estimate the required laser-injection power of each wavelength for satisfying the input DR and signal-to-noise ratio of O/E conversion as discussed so far. That is, each laser-injection power (P_λ) must confront log2(d) stages Y-junctions (3.01-dB power splitting and 0.05-dB loss per stage) and three MRM transmission DR losses (at least 2.5-dB loss per e-MRM$_i$, z-MRM$_i$, and y-MRM$_{ij}$ based on Fig. 4). Meanwhile, the aggregate absorption of all wavelength powers per PD needs to be confined within the linear DR of the O/E conversion circuit. The criterion for the maximum laser-injection power per wavelength can be expressed as follows:

$$P_{\lambda,max}\cdot\left(10^{\frac{-2.5\cdot dB}{10}}\right)^3\cdot\left(\frac{1}{d}\cdot 10^{\frac{-log_2(d)\cdot 0.05\cdot dB}{10}}\right)\cdot d\le DR_{O/E} \tag{15}$$

where $P_{\lambda,max}$ is not a strong function of "d" because $DR_{O/E}$ stays constant regardless of "d", and only the number of the Y-junction stages would gradually increase the power loss within each MVM row. For the case of $d = 32$ and $DR_{O/E} = 670\ \mu$W, $P_{\lambda,max}$ is about 4 mW; the total 32 laser-injection power (P_{inj}) for this WDM-based MVM operation is 128 mW, which however is directly proportional to the number of WDM wavelengths "d".

G. Thermal Tuning & Post-Fabrication Trimming

The high thermo-optic coefficient (1.86×10^{-4}1/K) of silicon [41] makes SiPh devices extremely sensitive to temperature variations; therefore, a proper thermo-control mechanism is necessary to stabilize the environment temperature for maintaining consistent characteristics of SiPh devices. On the other hand, this high-temperature sensitivity also helps to enable high-resolution SiPh characteristic tunability beyond the achievable resolution of the fabrication process especially for the WDM-based MVM; e.g., the 32 different radii of all M-MRM$_{ij}$ listed in Table I are distributed from 4.63 μm to 4.76 μm, which are impossible to be explicitly fabricated merely relying on the mask resolution of 45SPCLO. In CMOS-compatible SiPh process technology, tungsten heaters have been widely used for thermo-optic tuning [42], [43], and, according to the simulation result in 45SPCLO, a tungsten heater power efficiency per MRM can reach about 4.6 mW/$\Delta\lambda_{FSR}$. Therefore, if a d-by-d MPE-MVM accelerator contains "$d\cdot(2 + d)$" MRMs within a 1-mm^2 silicon area, the tuning ranges of each MRM need to cover about one $\Delta\lambda_{WDM} \approx \Delta\lambda_{FSR}/d \approx \Delta\lambda_{FSR,RTM}$, so the total heater power per d-by-d MVM is about $d\cdot(2 + d)\cdot$(4.6-mW)/d, which is 156 mW when $d = 32$.

Meanwhile, to accommodate a large number of MRMs for high-dimensional MVM in the attention-head, one heater source per MRM is impractical. Therefore, this paper proposed a hybrid tuning approach by combining the tungsten-heater approach for global coarse-tuning with the post-fabrication-trimming approach [44] for individual fine-tuning to cover the designated WDM spectrum as well as the PVT variations. For the example shown in Fig. 1, all MRMs in the entire MVM are partitioned into two tungsten-heater regions with two corresponding heater sources, and the area of each region shall be less than 1 mm^2 for confining random variation within one standard deviation. Thus, when one of the heaters is heating up its own region, all transmission responses of the MRMs belonging to this region are shifted together for globally tuning the resonance wavelengths into the fine-tuning ranges of the post-fabrication trimming mechanism. The post-fabrication trimming mechanism is basically realized by implanting a section of SOI rib waveguide with Ge through a photoresist mask on the top of each MRM cavity, so each MRM resonance wavelength can be trimmed by injecting a voltage pulse to anneal this Ge rib waveguide. This annealing calibration process can be done for each MRM individually as shown in Fig. 2, and the annealing pulse width for the targeted wavelength of each MRM is reached by the iterative feedback mechanism of the MRM transmission power received by its ADC to adjust the annealing pulse generator output pulse width until converging to the targeted wavelength [44], which is similar to the calibration process of finding e_i<2:0> for the comb-line power equalization through e-MRM$_i$. Although this post-fabrication trimming approach is tedious, it is thorough, reprogrammable, hardware reusable, and only consumes calibration time and power with almost zero overhead during regular MPE-MVM operations.

V. MPE Matrix-Matrix Multiplications

To perform an MPE-MMM operation, e.g., $Y_{d\times d}\cdot Z_{d\times n} = YZ_{d\times n}$, the input d-by-n matrix $Z_{d\times n}$ with elements denoted by z_{ij}, $i = 1\sim d$, $j = 1\sim n$, can be essentially split into "n" d-by-1 column vectors as shown in Fig. 8, and each column vector individually performs MVM with the matrix $Y_{d\times d}$ in an MPE-MVM unit and produce its own output column vector. After combining total "n" output column vectors from "n" parallelized MPE-MVMs, the outcome d-by-n matrix $YZ_{d\times n}$ with elements denoted by yz_{ij}, $i = 1\sim d$, $j = 1\sim n$, of the MMM operation can be obtained. Alternatively, this space-parallelism approach can be implemented by a time-multiplexing approach; e.g., a single MPE-MVM can process one of the input column

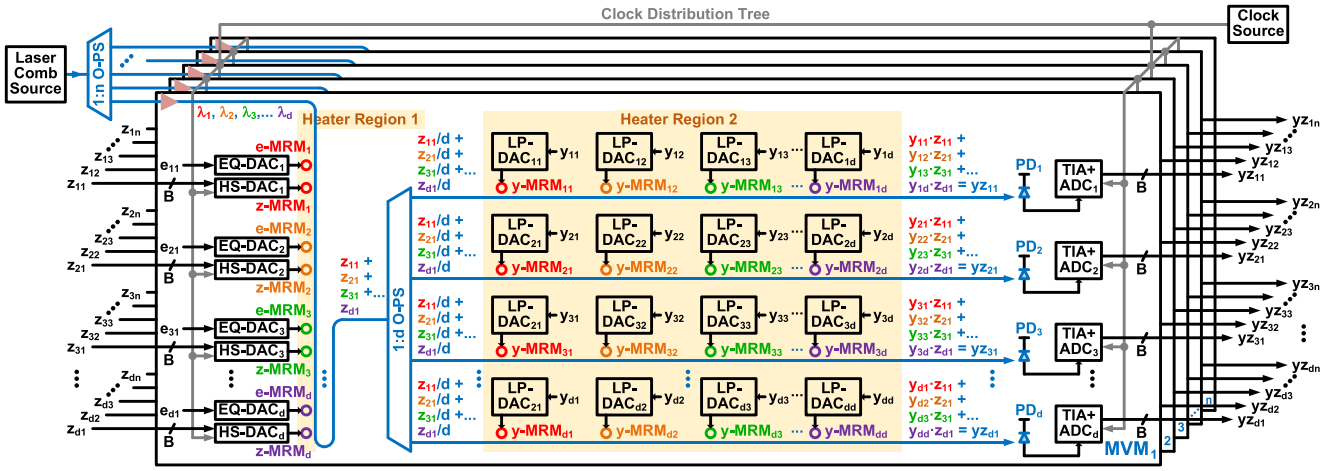

Fig. 8. The system block diagram of the MPE-MMM linear-algebra accelerator implemented by the MVM space-parallelism approach. Note that each "yz" presents a "single-word" variable, NOT "y times z".

vectors per iteration period, and electronic registers after the ADCs can collect all the output column vectors after "n" iteration cycles to form the final outcome matrix $YZ_{d\times n}$ per MMM operation. These two approaches linearly consume area and time with the column-dimension "n" of $Z_{d\times n}$, respectively. Therefore, an optimized power/area/throughput tradeoff can be done by mixing partial space-parallelism and partial time-multiplexing approaches based on realistic applications.

It is important to note that both space-parallelism and time-multiplexing approaches can be exploited to implement even higher dimensional MPE-MVMs by decomposing the vector and matrix elements into smaller dimensional MPE-MVMs. For the example of $d = 256$, a "$Y_{256\times256}\cdot Z_{256\times1}$" MPE-MVM can be realized by either executing 64 "$Y_{32\times32}\cdot Z_{32\times1}$" MPE-MVMs simultaneously (i.e., space-parallelism) or 64 times of a "$Y_{32\times32}\cdot Z_{32\times1}$" MPE-MVM (i.e., time-multiplexing) with additional power/area overhead from electronic digital summations to form the final result of $YZ_{256\times1}$. Either approach can avoid the requirements of generating a large number of comb lines ($= 256$) from the external optical comb source and a large $\Delta\lambda_{FSR}$ ($> 256\cdot\Delta\lambda_{WDM}$) per MRM with an unreasonably small radius, which can cause intolerable radiative losses and manufacturing random variations.

VI. MPE Double Matrix-Matrix Multiplications

As elaborated by (3) and (6), the double matrix-matrix multiplications (D-MMM) play an important role in the attention-head computations; e.g., $X_{n\times d}\cdot Y_{d\times d}\cdot Z_{d\times n} = XYZ_{n\times n}$ with elements denoted by xyz_{ij}, $i = 1 \sim n$, $j = 1 \sim n$ (note that each "xyz" presents a "single-word" variable, NOT "x times y times z"). Intuitively, two MPE-MMM units shown in Fig. 8 can be cascaded so that the first stage completes the operation of $Y_{d\times d}\cdot Z_{d\times n} = YZ_{d\times n}$ and then the second stage can do the $X_{n\times d}\cdot YZ_{d\times n} = XYZ_{n\times n}$. However, this consecutive MPE-MMMs architecture for the D-MMM functionality requires the intermediate multiplication result $YZ_{d\times n}$ to be converted from the photonic to the electronic domain and then back to the photonic domain, which is a d-by-n ADC/DAC power-and-area dominant O/E/O interface with no contribution to overall computation throughput and exposing the major downside of photonic computing as mentioned in Section I.

This paper proposed a single MPE-D-MMM unit to perform energy-efficient two consecutive MMM operations in the photonic domain together, which can completely eliminate the intermediate O/E/O overhead. As shown in Fig. 9, the overall D-MMM is established by "n" double matrix-vector multiplication (D-MVM) units in parallel, and each D-MVM unit performs two consecutive MVMs at once in the photonic domain. For example, the input vector $Z_{d\times1,j}$ of D-MVM$_j$ is one of the column vectors of the input matrix $Z_{d\times n}$. After converting this input vector $Z_{d\times1,j}$ into the photonic domain, $Z_{d\times1,j}$ is firstly multiplied by the MRM array of the matrix $Y_{d\times d}$, and then the first-stage MVM output column vector $YZ_{d\times1,j}$ ($= Y_{d\times d}\cdot Z_{d\times1,j}$) is generated. So far, the signal flow in Fig. 9 of D-MVM$_j$ completing the first-stage MVM operation is exactly the same as a single MVM operation in Figs. 1 and 8. Then, each element of $YZ_{d\times1,j}$ is duplicated by a 1-to-n O-PS for the second-stage MVM operation, i.e., $X_{n\times d}\cdot YZ_{d\times1,j}$. Note that each element of $YZ_{d\times1,j}$ already contains all wavelengths (λ_j, $j = 1 \sim d$) carrying their own weight factors. Therefore, to further perform MVM with $YZ_{d\times1,j}$, the control-voltage of each element (x_{ij}, $i = 1 \sim n$, $j = 1 \sim d$) of $X_{n\times d}$ requires a racetrack modulator (RTM) to influence the amplitudes of all wavelengths in each element of $YZ_{d\times1,j}$ all together; in other words, each RTM is driven by x_{ij} for a broadband (i.e., all WDM wavelengths) light-wave power modulation. Note that the detailed RTM design is elaborated in Section VI-A, and all x-RTM$_{ij}$, $i = 1 \sim n$, $j = 1 \sim d$, are essentially identical, so the indexes i and j for x-RTM$_{ij}$ are merely for distinguishing x-RTM$_{ij}$ driven by their pre-determined static digital x_{ij} data elements through their LP-DACs.

After the broadband modulations for element-to-element products of $X_{n\times d}$ and $YZ_{d\times1,j}$, the following "n" d-to-1 parallel linear Ge PDs are required to perform the summations of element-to-element products. As shown in Fig. 9, the photocurrent summations based on Kirchhoff's Current Law (KCL)

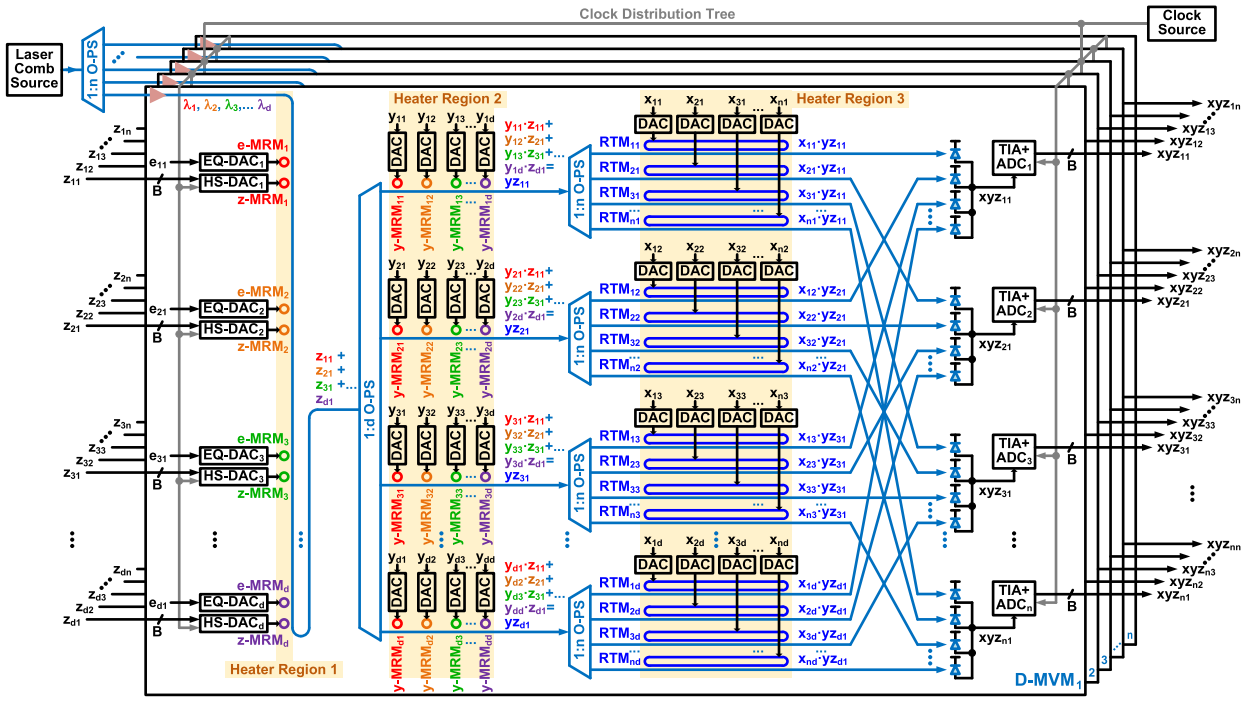

Fig. 9. The system block diagram of the MPE-D-MMM linear-algebra accelerator implemented by the D-MVM space-parallelism approach. Note that each "yz" presents a "single-word" variable, NOT "y times z"; each "xyz" also presents a "single-word" variable, NOT "x times y times z". Also, the PDs are actually distributed with the waveguides, and the signal crossovers are done by the PD electrical outputs with eight back-end metal layers in 45SPCLO.

of these "n" d-to-1 PDs represent the outcome vector of the D-MVM operation, $XYZ_{n\times1,j}$ ($=X_{n\times d}\cdot YZ_{d\times1,j}$). Note that the waveguides in 45SPCLO are formed by crystalline silicon in a single front-end layer, so the crossovers among the waveguides to reach the parallel PDs shown in Fig. 9 are for illustration purposes. In the realistic mask layout, the PDs are distributed with the waveguides, and signal crossovers are actually done by eight back-end metal layers in 45SPCLO carrying the electrical xyz_{ij} $i = 1 \sim n$, $j = 1 \sim n$, photo-currents from the outputs of the PDs merging at the inputs of the TIAs.

Overall, with either the space-parallelism approach in Fig. 9 or an equivalent time-multiplexing approach, the "n" D-MVM units can accomplish the whole D-MMM operation ($X_{n\times d}\cdot Y_{d\times d}\cdot Z_{d\times n} = XYZ_{n\times n}$) fully in the photonic domain for eliminating the intermediate O/E/O overhead and maintaining negligible computation latency and power consumption with extra area consumption and calibration effort due to additional RTMs, LP-DACs, O-PSs and PDs per D-MVM unit.

A. *Broadband Racetrack Modulators*

In the MPE-D-MMM described in this Section, the second multiplication in the photonic domain requires a broadband modulator to convert the second digital multiplicand x_{ij}, $i = 1 \sim n$, $j = 1 \sim d$, into consistent power-gains across all WDM wavelengths λ_{ij}, $j = 1 \sim d$, as shown in the bottom of Fig. 3(b). Note that the indexes i and j of x_{ij} and x-RTM_{ij} are only used for labeling the data elements of the multiplicand matrix $X_{n\times d}$, and they are not related to the indexes of the WDM wavelengths λ_i modulated by e-MRM_i/z-MRM_i, $i = 1 \sim d$, and λ_{ij} modulated by y-MRM_{ij}, $j = 1 \sim d$.

This paper proposed an optical comb-based racetrack modulator (RTM) [45] as shown in Fig. 3(b) to simultaneously modulate the light-wave powers across all WDM wavelengths in the MPE-D-MVM and MPE-D-MMM. The fundamentals of micro-ring and racetrack modulators are basically identical, but the dimensions of their resonance cavity are quite different. If the waveguide widths, gaps, and coupling lengths of the MRM and RTM are unified, then their resonance cavity lengths, L_{ij} and L_{RTM}, are the primary design parameters determining their power transmission responses. Under the resonance condition, the cavity of x-RTM_{ij}, $i = 1 \sim n$, $j = 1 \sim d$, respectively, in Fig. 9 simultaneously establishes constructive inferences with all wavelengths λ_{ij}, $j = 1 \sim d$, so that each first MVM dot-product result carried by all λ_{ij} light-wave powers in each straight waveguide can be altogether absorbed in the x-RTM_{ij} cavity, which is corresponding to a zero multiplicand, $x_{ij} = 0$. That is, designing an RTM with a power transmission spectrum matched with the aggregate power transmission spectrums of all the MRMs as the relation shown in Fig. 3(a) can perform an optical comb-based broadband modulation according to the P/N junction control voltage of x-RTM_{ij} based on x_{ij}.

To design x-RTM_{ij} having a consistent power transmission gain for all WDM wavelengths as shown in the bottom of Fig. 3(a), the perimeter L_{RTM} of x-RTM_{ij} for the whole WDM spectrum shall at least satisfy two criteria. First, the FSR $\Delta\lambda_{FSR\text{-}RTM,ij}$ of each resonance wavelength λ_{ij} determines each WDM isolation spacing $\Delta\lambda_{WDM,ij}$ [32]. Second, each resonance wavelength λ_{ij} under a particular resonance mode

TABLE II
PERFORMANCE SIMULATION RESULTS OF MPE-MVMS

	Vector Dimen.	Laser Inj. Power	Heater Power	SoC Power*	SoC Area	Data Precision	Clock Rate	Computation Throughput		Computation Density	Energy Consumption*
MPE-MVM in 45-nm Monolithic SiPh [This Work]	d	P_{inj} (mW)	P_{heat} (mW)	P_{SoC} (mW)	A_{SoC} (mm^2)	B (bit)	f_{CLK} (GHz)	$2{\cdot}d^2{\cdot}f_{CLK}$ (TOPS/s)	$d^2{\cdot}f_{CLK}$ (TMAC/s)	$d^2{\cdot}f_{CLK}/A_{SoC}$ (TMAC/s/mm^2)	$P_{SoC}/(d^2{\cdot}f_{CLK})$ (fJ/MAC)
	8	32	40.8	99.6	0.10	4	2	0.256	0.128	1.26	777.8
	16	64	79.2	198.7	0.33	4	2	1.024	0.512	1.56	388.0
	32	128	156.0	400.7	1.14	4	2	4.096	2.048	1.80	195.6
	64	256	309.6	818.0	4.16	4	2	16.384	8.192	1.97	99.8
	128	512	616.8	1701.1	15.77	4	2	65.536	32.768	2.08	51.9
	256	1024	1231.2	3653.3	61.12	4	2	262.144	131.072	2.14	27.9
ASIC-MVM Google TPU in CMOS	Vector Dimen.	SoC Idle Power		SoC Busy Power†	SoC Area	Data Precision	Clock Rate	Computation Throughput		Computation Density	Energy Consumption†
	d	P_{idle} (mW)		P_{SoC} (mW)	A_{SoC} (mm^2)	B (bit)	f_{CLK} (GHz)	$2{\cdot}d^2{\cdot}f_{CLK}$ (TOPS/s)	$d^2{\cdot}f_{CLK}$ (TMAC/s)	$d^2{\cdot}f_{CLK}/A_{SoC}$ (TMAC/s/mm^2)	$P_{SoC}/(d^2{\cdot}f_{CLK})$ (fJ/MAC)
v1 28-nm [47]	256	28000		40000	331	8	0.7	91.76	45.88	0.14	871.8
v4 7-nm [48]	256	55000		78571‡	400	8	1.05	137.62	68.81	0.17	1141.9

* Without including the external laser comb source power, P_{SoC} covers all photonic/electronic devices/circuits power, heater power (P_{heat}), and laser injection power (P_{inj}) on the single monolithic SiPh chip.
† Including all electronic digital circuits power consumption on the single CMOS chip in the Busy-mode.
‡ The Busy-mode power of TPUv4 is estimated by its Idle-mode power (= 55 W) and Busy-vs.-Idle power ratio of TPUv1 (= 1.43).

$m_{RTM,ij}$ (an integer) corresponds to its effective refractive index $n_{eff}(\lambda_{ij})$, silicon propagation constant $\beta(\lambda_{ij})$, and the common perimeter $L_{RTM} = 2\pi{\cdot}r_{RTM} + 2{\cdot}s_{RTM}$ [33]. These two criteria are summarized in (16) and (17), respectively, as follows:

$$\Delta\lambda_{FSR\text{-}RTM,ij} = \frac{\lambda_{ij}^2}{n_g(\lambda_{ij}) \cdot L_{RTM}} = \Delta\lambda_{WDM,ij}$$

$$\Rightarrow L_{RTM} = \frac{\lambda_{ij}^2}{n_g(\lambda_{ij}) \cdot \Delta\lambda_{WDM,ij}} \quad (16)$$

$$2\pi \cdot m_{RTM,ij} = \beta(\lambda_{ij}) \cdot L_{RTM} = \frac{2\pi}{\lambda_{ij}} \cdot n_{eff}(\lambda_{ij}) \cdot L_{RTM}$$

$$\Rightarrow L_{RTM} = \frac{\lambda_{ij}}{n_{eff}(\lambda_{ij})} \cdot m_{RTM,ij} \quad (17)$$

To satisfy both (16) and (17), all resonance wavelengths λ_{ij}, $j = 1 \sim d$, of the entire WDM spectrum are determined by the resonance mode requirements of x-RTM_{ij} as shown as follows:

$$m_{RTM,ij} = \frac{n_{eff}(\lambda_{ij})}{n_g(\lambda_{ij})} \cdot \frac{\lambda_{ij}}{\Delta\lambda_{WDM,ij}}$$

$$= \frac{n_{eff}(\lambda_{ij})}{n_{eff}(\lambda_{ij}) - \lambda_{ij} \cdot \frac{\partial n_{eff}(\lambda_{ij})}{\partial \lambda_{ij}}} \cdot \frac{\lambda_{ij}}{\Delta\lambda_{WDM,ij}} \quad (18)$$

Note that each $m_{RTM,ij}$ has to be a unique integer for $j = 1 \sim d$ since all constructive interferences simultaneously occur in the same cavity, L_{RTM}. In other words, the only way to set distinguishable "d" resonance wavelengths λ_{ij} by (17) with a common L_{RTM} is to assign "d" individual $m_{RTM,ij}$. In summary, the value of "d" with the requirements in (16), (17), and (18) determines the distribution of λ_{ij} and $\Delta\lambda_{WDM,ij}$ of the entire WDM wavelength spectrum for the MVM operation; then the laser wavelengths and the radii of all z-MRM_i and y-MRM_{ij} described in Section IV-D shall be designed accordingly to match the power transmission spectrums as shown in Fig. 3(a).

By using the same example in Section IV-D with a few iterations, the consecutive integer $m_{RTM,ij}$, $j = 1 \sim 32$, are chosen from 2321 down to 2290 so that λ_{ij} and $\Delta\lambda_{WDM,ij}$, $j = 1 \sim 32$, can be determined to satisfy the targeted 0.5-nm WDM isolation spacing with $L_{RTM} = 951.32$ μm as summarized in Table I. About the area of x-RTM_{ij}, if the radius r_{RTM} of the left/right-end half-circles is set to 5 μm, then the length s_{RTM} of the top/bottom straight waveguides is 459.95 μm. Therefore, including the primary racetrack resonator and peripheral keep-out halo, the silicon area per RTM is a 480-μm $\times$ 20-μm tile.

VII. PERFORMANCE SUMMARY AND CONCLUSION

The primary performance metrics of linear-algebra computing systems include computation throughput (TMAC/s), computation density (TMAC/s/mm^2), and energy consumption (fJ/MAC) [15], [19], [24]. Based on Section IV and Fig. 1, the required counts of the building blocks in a completely on-chip MPE-MVM can be scalable and parameterized into functions of "d" (proportional to "d" or "d^2"). After consolidating the building-block powers, areas, and bandwidths reported in Section IV, which were simulated by the Cadence design environment [46] with the GlobalFoundries 45SPCLO PDK, into this scalable architecture, the total power and area of the MVM accelerator can be calculated by summing the building-block powers and areas multiplied by their corresponding counts parameterized by "d" with additional margins for on-chip signal/clock routing, buffering, supply/bias distribution, and 45SPCLO design-rule-check (DRC) compliance.

According to the approach described above, the detailed power/area breakdowns and computing performance metrics of the MPE-MVM accelerator with its entire single-chip hardware and standalone MVM functionality are summarized in Table II to practically cover the overhead of the MPE integrations, conversions, and calibrations, including all electronic/photonic SoC building blocks with complete DACs/ADCs, on-chip digital interface, clock distribution, on-chip calibration hardware, laser injection power, and heater power, instead of only considering the power/area of the photonic devices in the performance evaluations [15], [19], [24]. For the same reason of thoroughness in practical hardware realization, the performance metrics of Google Tensor Processing Units (TPU) reported in [47], [48] are listed in Table II as well for comparison purposes since these TPUs are also fully integrated SoCs possessing complete MVM functionalities implemented by digital application-specific integrated-circuits (ASIC). Note that the performance metrics of the MPE-MMM and MPE-D-MMM accelerators can

be reasonably simulated and calculated according to those of the MPE-MVM accelerator because of the dimension scalability, space-parallelism, and time-multiplexing approaches described in Sections V and VI.

The performance scalability with the dimensional parameter "d" of the MPE-MVM accelerator listed in Table II shows that the power/area overhead of electronic circuits in the MPE-MVM is getting leveraged by the negligible photonic computing latency and power consumption when "d" is scaling up. In the cases of "d" ≥ 8, the MPE-MVM accelerator outperforms the ASIC counterparts in both computation density and energy consumption (the farthest right two columns of Table II). In particular, with the future advances in scaling the optical comb-line generations up to "d" $= 256$, the MPE-MVM accelerator can exhibit about $12.6\times$ computation-density and $40.9\times$ energy-efficiency superiority over the advanced ASIC-MVM accelerator (TPUv4) with the downside of lower data precision due to the "analog" computing in photonics. Finally, it is important to note that many novel SiPh devices and circuits are still being discovered and engineered for future foundry manufacturing; potentially, the performance metrics of the MPE accelerators can be further improved and scaled with the development of next-generation SiPh process technology.

REFERENCES

[1] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[2] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[3] "ChatGPT sets record for fastest-growing user base - analyst note," Reuters, Feb. 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[4] "How to use ChatGPT to write code," Pluralsight, Mar. 2023. [Online]. Available: https://www.pluralsight.com/blog/software-development/how-use-chatgpt-programming-coding

[5] "Latest version of ChatGPT aces bar exam with score nearing 90th percentile," ABA Journal, Mar. 2023. [Online]. Available: https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile

[6] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, Jul. 2021.

[7] P. Moorhead, "Silicon photonics are here and global foundries is innovating," Moor Insights & Strategy, Oct. 2022. [Online]. Available: https://gf.com/wp-content/uploads/2022/12/Silicon-Photonics-Are-Here-And-Global-Foundries-Is-Innovating-Final-V10.24.2022-2.pdf

[8] M. Rakowski et al., "45nm CMOS-silicon photonics monolithic technology (45CLO) for next-generation, low power and high speed optical interconnects," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, San Diego, CA, USA, 2020, pp. 1–3.

[9] C. Levy et al., "A 3D-integrated $8\lambda \times 32$ Gbps λ silicon photonic microring-based DWDM transmitter," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, San Antonio, TX, USA, 2023, pp. 1–2.

[10] H. Li et al., "A 112 Gb/s PAM4 silicon photonics transmitter with microring modulator and CMOS driver," *J. Lightw. Technol.*, vol. 38, no. 1, pp. 131–138, Jan. 2020.

[11] H. Li, G. Balamurugan, J. Jaussi, and B. Casper, "A 112 Gb/s PAM4 linear TIA with 0.96 pJ/bit energy efficiency in 28 nm CMOS," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Dresden, Germany, 2018, pp. 238–241.

[12] C. Sun et al., "A 45 nm CMOS-SOI monolithic photonics platform with bit-statistics-based resonant microring thermal tuning," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 893–907, Apr. 2016.

[13] N. Mehta, S. Buchbinder, and V. Stojanović, "Design and characterization of monolithic microring resonator based photodetector in 45 nm SOI CMOS," in *Proc. IEEE Eur. Solid-State Device Res. Conf. (ESSDERC)*, Cracow, Poland, 2019, pp. 206–209.

[14] N. Mehta et al., "A 12 Gb/s, 8.6μApp input sensitivity, monolithic-integrated fully differential optical receiver in CMOS 45 nm SOI process," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Lausanne, Switzerland, 2016, pp. 491–494.

[15] M. A. Nahmias et al., "Photonic multiply-accumulate operations for neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7701518.

[16] K. Kikuchi, "Fundamentals of coherent optical fiber communications," *J. Lightw. Technol.*, vol. 34, no. 1, pp. 157–179, Jan. 2016.

[17] C. Dragone, "Efficient N∗N star couplers using fourier optics," *J. Lightw. Technol.*, vol. 7, no. 3, pp. 479–489, Mar. 1989.

[18] R. A. Athale and W. C. Collins, "Optical matrix–matrix multiplier based on outer product decomposition," *Appl. Opt.*, vol. 21, no. 12, pp. 2089–2090, Jun. 1982.

[19] H. Zhou et al., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light: Sci. Appl.*, vol. 11, no. 1, Feb. 2022, Art. no. 30.

[20] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, pp. 52–58, Jan. 2021.

[21] L. Yang, L. Zhang, and R. Ji, "On-chip optical matrix-vector multiplier for parallel computation," *Proc. SPIE*, vol. 8855, pp. 100–104, 2013, doi: 10.1117/2.1201306.004932.

[22] D. Dang, B. Lin, and D. Sahoo, "LiteCON: An all-photonic neuromorphic accelerator for energy-efficient deep learning," *ACM Trans. Archit. Code Optim.*, vol. 19, no. 3, pp. 1–22, Aug. 2022.

[23] M. Li and Y. Wang, "An energy-efficient silicon photonic-assisted deep learning accelerator for Big Data," in *Proc. Conf. Wireless Commun. Mobile Comput.*, 2020, vol. 2020, no. 1, pp. 1–11.

[24] C. Huang et al., "Prospects and applications of photonic neural networks," *Adv. Phys.: X*, vol. 7, no. 1, pp. 1–63, 2022.

[25] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, pp. 102–114, Jan. 2021.

[26] H. Hu and L. K. Oxenløwe, "Chip-based optical frequency combs for high-capacity optical communications," *Nanophotonics*, vol. 10, no. 5, pp. 1367–1385, 2021.

[27] R. Hu, B. Tian, S. Yin, and S. Wei, "Efficient hardware architecture of softmax layer in deep neural network," in *Proc. IEEE 23rd Int. Conf. Digit. Signal Process. (DSP)*, Shanghai, China, 2018, pp. 1–5.

[28] X. Yang, B. Yan, H. Li, and Y. Chen, "ReTransformer: ReRAM-based processing-in-memory architecture for transformer acceleration," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, San Diego, CA, USA, 2020, pp. 1–9.

[29] T.-C. Hsueh et al., "A 25.6 Gb/s differential and DDR4/GDDR5 dual-mode transmitter with digital clock calibration in 22 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, USA, 2014, pp. 444–445.

[30] D. L. Feucht, *Handbook of Analog Circuit Design*. San Diego, CA, USA: Academic, 1990.

[31] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, 3rd ed., New York, NY, USA: Oxford, 2011.

[32] E. Hecht, *Optics*, 5th ed. Upper Saddle River, NJ, USA: Pearson, 2015.

[33] T. Hansson, D. Modotto, and S. Wabnitz, "Analytical approach to the design of microring resonators for nonlinear four-wave mixing applications," *J. Opt. Soc. Amer. B*, vol. 31, no. 5, pp. 1109–1117, 2014.

[34] J. Singh et al., "Neuromorphic photonic circuit modeling in Verilog-A," *APL Photon.*, vol. 7, no. 4, pp. 1–15, Apr. 2022.

[35] "INL/DNL measurements for high-speed analog-to-digital converters (ADCs)," Analog Devices Technical Articles, Nov. 2001. [Online]. Available: https://www.analog.com/en/resources/technical-articles/inldnl-measurements-for-types-of-highspeed-adcs.html

[36] E. Sackinger, *Analysis and Design of Transimpedance Amplifiers for Optical Receivers*. Hoboken, NJ, USA: Wiley, 2018.

[37] B. Razavi, *Design of Integrated Circuits for Optical Communications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2012.

[38] J. Kim, B. S. Leibowitz, J. Ren, and C. J. Madden, "Simulation and analysis of random decision errors in clocked comparators," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 56, no. 8, pp. 1844–1857, Aug. 2009.

[39] C. Ozcan, M. Mojahedi, and J. S. Aitchison, "Short, broadband, and polarization-insensitive adiabatic Y-junction power splitters," *Opt. Lett.*, vol. 48, no. 18, pp. 4901–4904, 2023.

[40] M. Glick et al., *Integrated Photonics for Data Communication Applications*. Amsterdam, The Netherlands: Elsevier, 2023.

[41] S. Y. Siew et al., "Review of silicon photonics technology and platform development," *J. Lightw. Technol.*, vol. 39, no. 13, pp. 4374–4389, Jul. 2021.

[42] A. Masood et al., "CMOS-compatible tungsten heaters for silicon photonic waveguides," in *Proc. IEEE Int. Conf. Group IV Photon. (GFP)*, San Diego, CA, USA, 2012, pp. 234–236.

[43] P. Dong et al., "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Opt. Exp.*, vol. 18, no. 19, pp. 20298–20304, 2010.
[44] H. Jayatilleka et al., "Post-fabrication trimming of silicon photonic ring resonators at wafer-scale," *J. Lightw. Technol.*, vol. 39, no. 15, pp. 5083–5088, Aug. 2021.
[45] C. Arlotti, O. Gauthier-Lafaye, A. Monmayrant, and S. Calvez, "Achromatic critically coupled racetrack resonators," *J. Opt. Soc. Amer. B*, vol. 34, no. 11, pp. 2343–2351, 2017.
[46] Cadence Design Systems, 1990. [Online]. Available: https://www.cadence.com/en_US/home.html
[47] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Archit. (ISCA)*, Toronto, ON, Canada, 2017, pp. 1–12.
[48] N. P. Jouppi et al., "Ten lessons from three generations shaped Google's TPUv4i," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Archit. (ISCA)*, Valencia, Spain, 2021, pp. 1–14.

Tzu-Chien Hsueh (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of California, Los Angeles, CA, USA, in 2010. From 2001 to 2006, he was a Mixed-Signal Circuit Design Engineer in Hsinchu, Taiwan. From 2010 to 2018, he was a Research Scientist with Intel Lab Signaling Research and an Analog Engineer with Intel I/O Circuit Technology, Hillsboro, OR, USA. Since 2018, he has been an Assistant Professor of electrical and computer engineering with the University of California, San Diego, CA, USA. His research interests include wireline electrical/optical transceivers, clock-and-data recovery, data-conversion circuits, on-chip performance measurements/analyzers, and digital/mixed signal processing techniques. He was the recipient of multiple Intel Division and Academy Awards from 2012 to 2018, 2015 IEEE Journal of Solid-State Circuits (JSSC) Best Paper Award, 2020 NSF CAREER Award, and 2022 UCSD Best Teacher Award. He served on the Patent Committee for Intel Intellectual Property (Intel IP) and Technical Committee for Intel Design & Test Technology Conference (DTTC) from 2016 to 2018. Since 2018, he has been on the Technical Program Committee for IEEE Custom Integrated Circuits Conference and the Guest Associate Editor for IEEE SOLID-STATE CIRCUITS LETTERS.

Yeshaiahu Fainman (Life Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the Technion-Israel Institute of Technology, Haifa, Israel, in 1979 and 1983, respectively. He is currently an inaugural ASML/Cymer Chair Professor of advanced optical technologies and a Distinguished Professor of electrical and computer engineering with the University of California, San Diego (UCSD), San Diego, CA, USA. He is directing research of the Ultrafast and Nanoscale Optics Group, UCSD and made significant contributions to near field optical phenomena, nanoscale science and engineering of ultra-small, sub-micrometer semiconductor light emitters and nanolasers, inhomogeneous and meta-materials, nanophotonics, and Si Photonics. He contributed more than 340 manuscripts in peer review journals and more than 560 conference presentations and conference proceedings. His research interests include field optical science and technology with Si Photonics applications to information technologies and biomedical sensing. He contributed to editorial and conference committee works of various scientific societies including IEEE, SPIE, and OPTICA. He is a Fellow of OPTICA (former OSA), IEEE, and SPIE. He was the recipient of the Miriam and Aharon Gutvirt Prize, Lady Davis Fellowship, Brown Award, Gabor Award, Emmett N. Leith Medal, Joseph Fraunhofer Award/Robert M. Burley Prize, and OPTICA Holonyak Award.

Bill Lin (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1985, 1988, and 1991, respectively. He is currently a Professor of electrical and computer engineering with the University of California, San Diego, CA, USA, where he is actively involved with the Center for Wireless Communications, Center for Networked Systems, and Qualcomm Institute in industry-sponsored research efforts. His research has led to more than 200 journal and conference publications, including a number of Best Paper awards and nominations. He also holds five awarded patents. He was the General Chair and on the executive and technical program committee of many IEEE and ACM conferences. He was an Associate and Guest Editors for several IEEE and ACM journals.