Optics EXPRESS

Foundry fabricated compact slow-light Mach-Zehnder modulator and photodetector for on-chip analog photonic computing

AMIR BEGOVIĆ,[1] MENG ZHANG,[1] DENNIS YIN,[2] NICHOLAS GANGI,[1] JIAQI GU,[2] AND Z. RENA HUANG[1,*]

[1] *Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[2] *School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA*
[*] *huangz3@rpi.edu*

Abstract: This work presents a scaling pathway of on-chip analog photonic computing using foundry-fabricated silicon electro-optic (EO) slow-light Mach-Zehnder modulators (SL-MZMs) and compact Ge photodetectors (PDs) to construct a computing unit. Two SL-MZMs with phase shifter (PS) lengths of 500 μm and 150 μm are studied in this work. The bit resolution, nonlinearity, clock frequency, and power consumption of the photonic computing link, including an RF amplifier, on-chip SL-MZM, and a PD, are thoroughly investigated. The computing link using the SL-MZM with 500 μm has demonstrated a low normalized mean square error (NMSE) of 0.0305 at 8-bit resolution under 3.2 GHz clock frequency. Under the setting of 6-bit resolution at a clock frequency of 800 MHz, high computing accuracy was achieved with a measured NMSE of 0.0018 using the SL-MZM with 150 μm PS length. Using the Google Speed Commands dataset to run a voice keyword spotting task, we determine that 6-bit resolution operating at 3.2 GHz achieves the optimal power-accuracy trade-off. We show a 20× improvement in energy efficiency and a 3.35× improvement in area efficiency compared to NVIDIA V100 GPU ["Volta: Performance and programmability," IEEE Micro **38**(2), 42 (2018)]. These results show that our compact SL-MZMs and PDs promise to scale up photonic computing for practical machine-learning applications.

1. Introduction

Photonic tensor cores have garnered significant interest recently due to their potential applications in machine learning (ML) and artificial intelligence (AI) [1–3]. Unlike electronic systems, photonic computing can handle vast amounts of data simultaneously due to its inherent parallelism and high bandwidth capabilities [4], offering unprecedented advantages in computing efficiency, power, and speed. Matrix-vector multiplication is one of the fundamental operations for any computing task, which can be readily realized using cascaded Mach-Zehnder modulators (MZMs) in the optical domain. For AI tasks, weights refreshed at slower speeds (~MHz) can be encoded on thermally tuned optical modulators. Input data, such as those collected from various sensors, often needs real-time signal processing, thus requiring a fast electro-optic (EO) modulator at hundreds of MHz to multiple GHz operation speed. These EO modulators in PTCs are critical components for dot-product computing as they determine both clock frequency and bit precision.

For PTC computing, a large array of EO modulators is needed to construct the matrices for the multiplication operation. Scaling the PTC becomes an increasingly challenging issue for the practical implementation of PTC on photonic chips. Integrated micro-ring resonator (MRR) modulators, with typical radii ~10 μm for o-band and c-band operation, are an appealing solution due to their compactness and versatile use as optical switches, add-drop components, filters, and more [5,6]. However, there are several practical challenges in utilizing a large MRR array for

#540194
Journal © 2024
https://doi.org/10.1364/OE.540194
Received 26 Aug 2024; revised 5 Oct 2024; accepted 10 Oct 2024; published 1 Nov 2024

optical accelerators. First, to minimize thermal cross-talk, the MRR spacing is often kept at around 60-100 μm, so the actual chip surface area consumption is much larger than the MRR size [7]. Second, It becomes increasingly difficult to simultaneously tune to maintain all individual MRR to their desirable resonant conditions for a large array with tens to hundreds of MRR [8].

To address this scaling challenge, using silicon EO MZMs, which have higher thermal robustness, has become an appealing solution. Gratings tend to have lower quality (Q) factor values than rings which makes them less sensitive to environmental factors. Dispersion-engineered SL photonic structures with constant group index n_g over a spectrum range have exhibited greater than 40°C of temperature stability [9]. However, conventional Silicon EO modulators with a dimension of 2 mm or longer are not suited for large array integration on Si photonic chips. In recent years, photonic crystal (PhC) and 1D Bragg grating (BG) based slow-light Mach-Zehnder modulators (SL-MZMs) with enhanced light-matter interaction to shorten the phase shifter (PS) length have been widely explored for data center or 5G/6G communication applications [9–13]. While these compact SL-MZMs have been explored extensively for communication applications [13], in this work, we extend the utilization of SL-MZMs for analog photonic computing with large array integration. The evaluation of an on-chip optical link, comprised of essential photonic components for computing, i.e., a SL-MZM connected with an on-chip Ge photodetector (PD), is the focus of this work and serves as the proof-of-concept of SL-MZM for PTC in time-multiplexed computing architecture [14].

Two 1D BG-assisted SL-MZMs are studied in this work, with PS lengths of 500 and 150 μm, the latter of which has the PS length reduced aggressively in favor of scaling. These short MZMs can only achieve a fraction of π phase shift, offering a trade-off of increased linearity and degraded signal-to-noise (SNR) ratio. The nonlinearity analysis and bit resolution calculation based on quantifying MSM response curve residuals of a full π range no longer applies [15]. Detailed noise analysis are conducted in this work that reveals that the noise in the RF amplifier and on-chip Ge PD are the major sources contributing to the computing error besides inherent quantization noise. As the clock rate exceeds 1 GHz, the rise and fall time, limited by the Optical DAC and PD response, also contribute to computing accuracy. Therefore, in this work, we also aim to develop a theoretical framework for using aggressively scaled MZMs for computing tasks with detailed analysis on bit resolution, NMSE, clock frequency, and power consumption and their relation with link linearity, SNR, rise/fall time, and compound noise of the analog computing link. To our knowledge, it is the first time that an on-chip compact SL-MZM with digital on-chip PDs has been explored for PTC computing applications.

The measured photonic link terminal characteristics for computing, including power, clock frequency, and chip surface area, are incorporated in a computing simulator for a Voice Keyword Spotting task built on a previously proposed tensor core architecture TeMPO [14]. Computing accuracy, chip area analysis, power simulation analysis, and computational efficiency are evaluated for the two MZMs, and an optimal solution studied is identified and compared to a state-of-the-art graphics processing unit (GPU).

2. On-chip SL-MZMs and photodetectors

To enable scaling of the PTC, it is pivotal to fabricate all components using foundry services for manufacture consistency. In this work, both the on-chip Si SL-MZM and Ge PD were fabricated at AIM Photonics foundry [16], riding on a multi-project wafer (MPW) run.

2.1. *Si SL-MZM devices design*

Si MZMs are common components in process design kits (PDKs) offered by all Si Photonics foundries. The standard MZM in the AIM PDK library features a PS length of 2 mm. A reduced PS length would lead to less than π phase shift under the same voltage swing, resulting in a lower extinction ratio. Using Bragg gratings, a slow-down factor of 3 to 4 is attainable near the photonic band gap [12]. We designed two SL-MZMs, one with a PS length of 500 μm, aiming for a similar extinction ratio of a regular MZM with a PS length of ~1.5-2 mm and an aggressively scaled-down SL-MZM with a very short PS of 150 μm. We refer to the first one as MZM A and the second one as MZM B.

The schematic of the SL-MZM photonic structures are shown in Fig. 1 where MZM A and B have similar grating perturbation structure but differ in geometry dimensions. The Y-splitter/combiner is inversely designed to reduce beam reflection. Both MZMs have inner waveguide width (W_{min}) of 400 nm. The grating area width (W_{max}) is 800 nm for MZM A and increased to 1.5 μm for MZM B to increase the slow light effect for higher modulation efficiency. Additionally, the doping concentration in the P region of MZM B is increased to $\sim 10^{19} cm^{-3}$ for stronger index modulation. The doping concentration and device dimension are labeled in the cross-sectional view in Fig. 1(b).

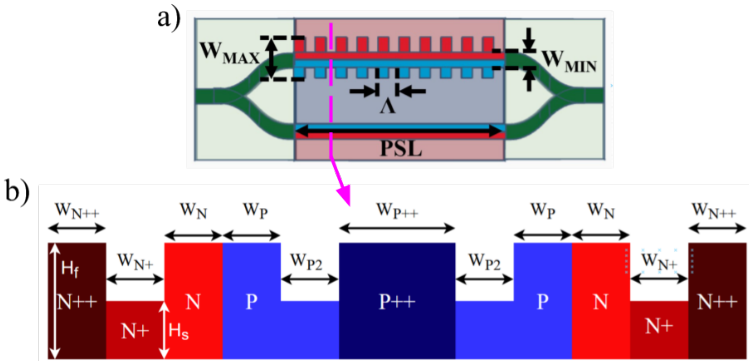


Fig. 1. a) Top view of the SL-MZMs where W_{max} = 800 nm, W_{min} = 400 nm, and grating period Λ= 290 nm. The pink dotted line and arrow indicate the cross section location shown in b). b) Cross-sectional view of the SL-MZMs with dimensions of H_f = 220 nm, H_s = 110 nm, W_{N++} = 4.75 μm, W_{N+} = 7.75 μm, W_N = 200 nm, W_P = 200 nm, W_{P2} = 2.5 μm, and W_{P++} = 4.5 μm (Not drawn to scale).

2.2. On-chip photodetectors

For optical computing, where both positive and negative values need to be represented, balanced PDs are commonly used where the current flow direction can indicate the sign of an intensity level [17]. Those two PDs in the balanced PD configuration are identical but placed in a series connection. To quantify the physical layer PD performance, we focus on a single PD connected to the SL-MZM, with analysis emphasizing PD responsivity, noise, sensitivity, and bandwidth.

The AIM PDK offers two c-band PD options, namely digital PD and analog PD, both fabricated on thin-film germanium in a waveguide coupled PiN structure [16]. The analog PD has a higher electrical bandwidth with over 30 times the chip space than the digital PD. For scaling consideration, we selected the digital PD in this work.

3. Experimental setup

The optical analog computing link performance was evaluated for vertical bit resolution under various clock frequencies from 100 MHz to 3.2 GHz. Multi-level signals were first generated digitally and then encoded electronically by an arbitrary wavefunction generator (AWG) with a maximum voltage output of 400 mV. The AWG signal was then amplified by an RF amplifier to produce a maximum of 2.5 V_{pp} amplitude. A schematic of the test setup is shown in Fig. 2. For simplicity, in the following chapter discussion, we use the term "analog computing link" to refer to the entire photonic circuit for testing, including AWG, RF amplifier, bias tee, SL-MZM, on-chip PD, and oscilloscope. The tunable laser has a maximum optical power of 13 dBm near λ = 1.55 μm, while the measured fiber-to-chip insertion loss is approximately 5 dB. The polarizer sets the optical mode to TE polarization.

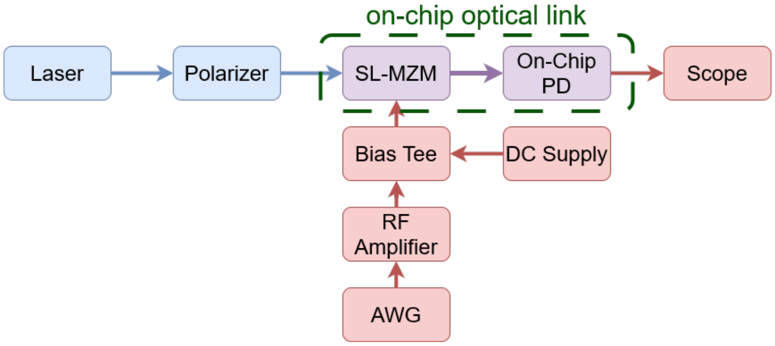


Fig. 2. Test setup diagram of the MZMs. Blue connections and boxes represent optical components, red ones represent electrical, and purple represents electro-optical. The green dotted box represents the portion of the test setup that is located on-chip and comprises the optical link.

Sufficient electrical bandwidth with a large dynamic range and good linearity are crucial in optical computing hardware testing. A summary of testing equipment used for bit resolution characterization is shown in Table 1. The limiting factor seems to be the AWG with a bandwidth of 12.5 GHz.

Table 1. Test equipment relevant performance parameters.

Testing equipment	Electrical bandwidth	Bit resolution
AWG (Tektronix 70002A)	12.5 GHz (Sampling rate: 25 GSa/s)	10-bit
RF amplifier (Picosecond 5867)	15 GHz	N/A
Bias Tee (Picosecond 5545)	20 GHz	N/A
GSG probes (Picosecond 40A)	40 GHz	N/A
SMA cable	18 GHz	N/A
Real-time oscilloscope (Keysight UXR0334B)	33 GHz (Sampling rate: 128 GSa/s)	10-bit

4. Results and discussion

4.1. DC characterization

We first performed a wavelength sweep to characterize the DC transmission spectrum of both MZMs, as shown in Fig. 3. The SL-MZMs operate at the edge of the photonic band gap, where a strong slow-light effect is present. Both MZM A and MZM B exhibit a measured photonic band gap of ~20 nm. Using the interferometric method [18], we obtained a maximum measured group index n_g of ~16. Two sinusoidal-shaped dips are detected right next to the photonic band gap due to destructive interference of the light in the two arms of the MZM. The full width half maximum (FWHM) of the two resonance dips are measured to be 11 nm, 1 nm respectively for MZM A and 4 nm, 1.1 nm respectively for MZM B.

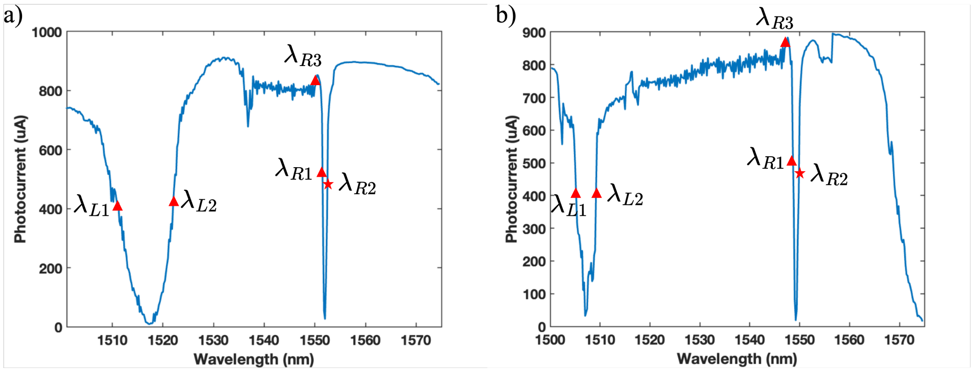


Fig. 3. DC transmission spectrum of a) MZM A and b) MZM B. 5 operation wavelengths (2 left to band gap and 3 right to bandgap) at different quadrature points are compared, and λ_{R2} (marked by red stars) turns out to perform the best for both MZMs and are selected to be the operation wavelengths.

For most Si MZMs, a thermal PS will be placed in both arms of the MZMs to adjust the quadrature points [13]. In this work, we scanned the wavelength to determine the optimal operation condition. Five potential operation wavelengths at the quadrature points are identified and marked in red in Fig. 3 for high linearity, extinction ratio (ER), and signal-to-noise ratio (SNR). Although R3 is also at the quadrature point, the propagation loss is high due to the increased group index near the photonic bandgap, resulting in reduced optical power. Testing shows that operation at R1 and R2 has produced similar ER while R2 point, i.e., 1551.2 nm for MZM A and 1548.6 nm for MZM B, gives the best computing accuracy due to the best linearity and are therefore selected to be the operation wavelengths for testing. MZMs A and B were able to be tested at R2 wavelengths without any thermal tuning or stabilization schemes to compensate for drift due to environmental conditions.

While the MZM PN junction is under reverse bias, the modulator operating voltage needs to be kept smaller than the breakdown (V_{br}). The measured V_{br} of MZM A and B are approximately 7.9 V and 6.5 V, respectively. MZM B has a higher doping concentration in the P region, so it has a smaller depletion region with a higher electric field, resulting in lowered breakdown voltage. MZM B with lower V_{br} also implies more efficient carrier plasma modulation. We searched the optimal DC bias voltage of the MZMs V_o over which a multi-bit level in a span of peak-to-peak voltage swing V_{pp} is set. For an ideal MZM subjected only to sinusoidal nonlinearity, the operation voltage should be chosen at the quadrature point. For silicon MZM with extra nonlinearity terms due to mode profile modulation, carrier injection, and alignment variation of the PN junction, the quadrature point is not necessarily the optimal bias point. In this work, we scanned the DC bias voltage in the range of 0 V to breakdown while monitoring the multiple bits quality displayed on the scope and ultimately set the V_o = -3 V for both MZMs.

For on-off keying (OOK) modulation in a communication channel, V_{pp} was superimposed on the DC bias V_o for bit rate testing, i.e., eye diagram measurement. For a computing task, V_{pp} marks the total voltage span of the bit levels while V_o refers to the middle setting of the V_{pp}. For OOK modulation in digital communication, nonlinearity isn't an issue so a higher V_{pp}, typically 5-7 V [19] is selected to obtain larger extinction ratio. However, higher driving voltages lead to greater nonlinearities along the sinusoidal curve, so performance is not necessarily better with higher V_{pp} for PTC applications. A driving voltage of 2.5 V_{pp} is chosen after analyzing the linearity of a staircase plot with multiple levels.

The MZM extinction ratio is characterized at the junction bias of 4.25 V and 1.75 V. The measured ER is 2.02 dB for MZM A and 1.17 dB for MZM B at their chosen operation wavelength. MZM B exhibits lower ER due to aggressively shrunken PS length. With V_{pp} at 6 V, the tested ER for MZM A is 4.6 dB, on par with non-slow light modulators with PS of 2 mm [19].

SL-MZMs are expected to have higher insertion loss when they operate in their slow-light spectrum due to enhanced light-matter interaction. At a wavelength of 1555 nm, where MZM A and B exhibit negligible slow-light effects, an insertion loss (IL) of about 1 dB was measured with both arms under no bias for both MZMs. Although MZM B is shorter, the higher doping concentration increases loss per unit length. At slow-light wavelengths of 1551.2 nm and 1548.6 nm, the tested IL increases to about 4 dB for MZMs A and B. Since neither MZM operates at the peak of the sinusoidal curve during slow-light operation, the reported IL here also included the optical power reduction due to the phase difference between the two arms. This level of loss is tolerable for such architectures that do not utilize cascading modulators [14] which highly increases accumulated insertion loss.

4.2. Nonlinearity characterization

The number of bit (M) levels scales with 2^M, so a larger linear response range is crucial for attaining higher bit precision in an optical link. The on-chip Ge PD can be assumed a linear component at low to moderate optical power levels so the RF amplifier and SL-MZM are the major components that cause nonlinearity. For all types of Mach-Zehnder modulators, the sinusoidal transfer function introduces an intrinsic first-order nonlinear term. Additionally, the Si phase shifter-based PN junction carrier plasma modulation introduces another nonlinearity factor [20]. The depletion width in a step junction follows a square root function with respect to the modulator voltage V, as shown in Eq. (1) [20]:

$$W_D = \sqrt{\frac{2\epsilon_{Si}}{q}\frac{N_D + N_P}{N_D N_P}(V_{bi} - V)} \tag{1}$$

where N_D and N_P are doping concentrations in the N and P regions, respectively, ϵ_{Si} is the permittivity of silicon, q is the elementary charge, and V_{bi} is the built-in voltage. In real devices, the doping profile of the junction often varies gradually, leading to a more complex response function. The effective index of the Si waveguide also varies nonlinearly with the partially depleted Si junction waveguide, and this complexity is further exemplified if the lateral junction is not perfectly aligned at the center of the Si waveguide and if the mode is not completely confined.

The phase modulation-induced nonlinearity of the MZM is characterized by the standard deviation (σ) of the residuals of the MZM power transmission curve to a linear curve within the entire multi-bit voltage range [21]. The bit precision B_{nonlin} due to MZM nonlinearity can be derived from

$$B_{nonlin} = \log_2(\frac{1}{\sigma}) \tag{2}$$

We characterized the dynamic nonlinear response of the MZM by driving it with a staircase signal train at 10-bit resolution (the highest setting of the AWG in our testing), corresponding to $2^{10} = 1024$ bit levels. The output of the AWG signals is fed to the RF amplifier so the MZM can be driven at the voltage range matching the bit precision testing. The 10-bit least significant bit (LSB) interval in the testing is 2.5 mV for both MZMs.

The normalized readout of the AWG, RF amplifier, and the on-chip optical link (MZM and PD) responses are plotted in Fig. 4. The reference signal in the graph refers to the digitally generated staircase signals, i.e., the ideal response curve to which the residuals are measured to compute σ and B_{nonlin}. During the experiment, the bit level holding time is set at 0.32 ns, corresponding to a total transition time of ~0.32 μs for scanning the entire 1024 levels. This transition time is significantly longer than the system rising/falling response time, therefore the response time induced signal distortion can be neglected in the staircase test. Details of rise and fall time testing results will be reported in section 4.4.

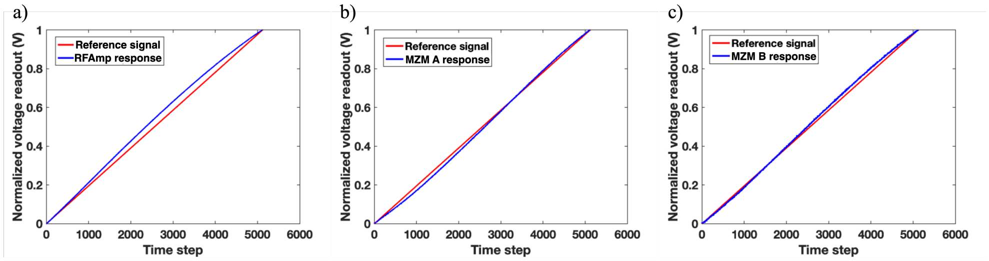


Fig. 4. Nonlinearity characterization of a) RF amplifier, b) MZM A, and c) MZM B. Testing was conducted at a clock frequency of 3.2 GHz with a sampling holding time of 0.3125 ns. The total transition time of each testing is 0.32 μs.

The measured standard deviation σ and computed B_{nonlin} results are summarized in Table 2. As indicated by Fig. 4, the nonlinearity of the RF amplifier results in a σ value of 0.0311, a 32-fold increase compared to that of the AWG. The σ of MZM A and B in both on-chip optical links exhibit lower values than that of the RF amplifier, indicating that the nonlinearity of the RF amplifier in the test acts to partially cancel out the nonlinearity of both MZMs. The technique of using a purposely designed nonlinear digital-to-analog converter (DAC) to drive MZM for nonlinearity compensation was reported in Z. Zhou et al. [22]. B_{nonlin} at 6-bit was obtained for both MZM A and B, higher than that of the RF amplifier.

Table 2. Nonlinearity Characterization of the MZMs.

Parameter	AWG	RF Amp	MZM A	MZM B
σ	0.0012	0.0311	0.0151	0.0141
B_{nonlin}	10	5	6	6

4.3. Noise limited vertical bit resolution

The accuracy of a real optical computing task can be limited by several factors, including noise effects, system nonlinearity, and system bandwidth, i.e., rise and fall time between discrete levels. In this section, we first study how the noise effects impact the bit resolutions by checking the staircase signal response of the system. Using staircase signals at a driving frequency of 200 MHz, we tested the optical system at 5-, 6- and 7-bit of resolution, illustrated in Fig. 5. Discrete voltage levels at 5- and 6-bit resolutions are distinguishable between neighbor levels while visually indistinguishable at 7-bit. We attribute the maximum attainable vertical bit resolution in this experiment to the SNR. Similar to digital communication systems where SNR correlates with bit error rate, in optical computing, the SNR impacts the computing accuracy.

The noise of the on-chip optical link consists of thermal and shot noise in the PD, noise from the RF amplifier, and the testing equipment. The RF amplifier has a listed noise figure (NF) of 5 dB on its specification sheet. We measured the noise floors of the RF amplifier, on-chip PD, and oscilloscope to quantify the relative weight in affecting the overall signal quality. The real-time scope exhibits a noise of ~0.45 mV under no external connection, while the noise floor on the scope increases to ~ 1 mV when the PD is under reverse bias via the GSG probing tip with laser off. The noise of the RF amplifier appears most detrimental to computing accuracy as it causes the optical signal intensity to fluctuate in the optical path. The recorded noise at the on-chip PD

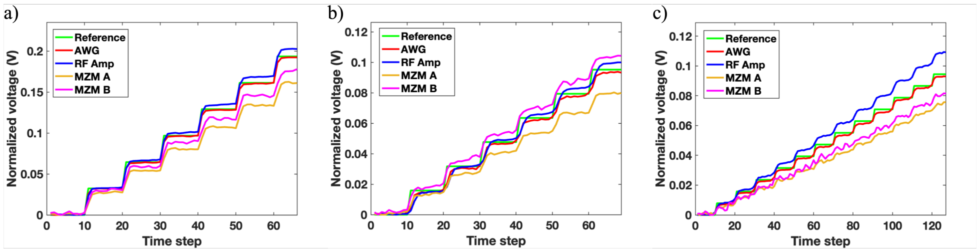


Fig. 5. Zoom-in plots of the system staircase signal responses at 200 MHz clock frequency with a) 5-bit, b) 6-bit and c) 7-bit vertical resolutions

is ~2.3 mV for MZM A and ~1.4 mV for MZM B, respectively. As MZM A has a longer PS length and greater E/O conversion efficiency, the noise from the RF amplifier noise is the major source, whereas, for MZM B, the testing equipment noise plays a major role compared to the RF amplifier. We anticipated an optimal MZM length, or E/O conversion efficiency, for a maximum SNR under the same laser output intensity.

4.4. Clock frequency and NMSE

The system bandwidth is another crucial factor that impacts computing accuracy, and the bandwidth-limited rise/fall time determines the clock frequency at which the system can operate. To quantify the impact from rise/fall time, we drive the RF amplifier and two MZMs separately with a 10 ns long square pulse with 2.5 V amplitude. The 10%-to-90% rise (fall) time is obtained to be 0.11 (0.13) ns, 0.83 (0.83) ns, and 0.70 (0.98) ns for the RF amplifier, MZM A and MZM B, respectively, as illustrated in Fig. 6.

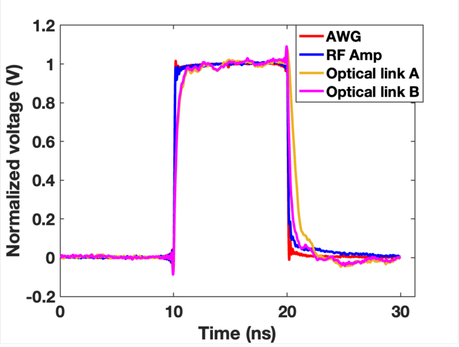
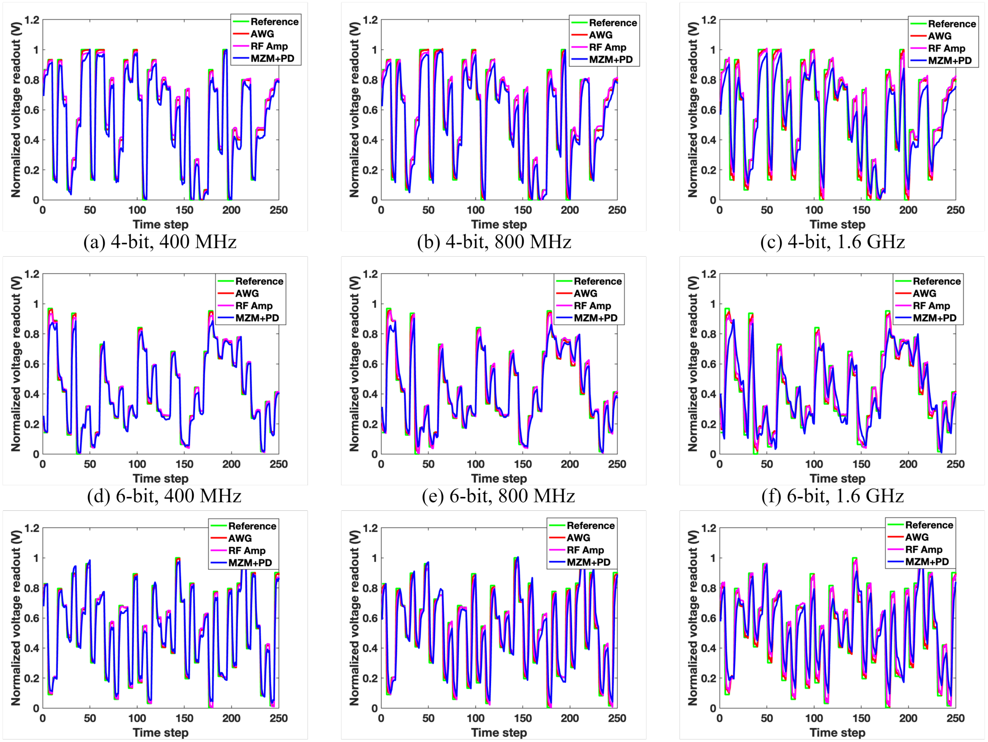


Fig. 6. Response time characterization. The pulse response of the RF amplifier (blue) includes the response time of AWG (red), and the pulse response for both optical links includes the response time of both AWG and the RF amplifier.

The transient response of the optical link to encoded signals with random distribution at different clock frequencies is analyzed next. The input data has an equal probability of being generated at any one of the 2^M levels for M-bit resolution. Examples of the traces for 4-, 6-, and 8-bit resolution testing are shown for MZMs A and B in Figs. 7 and 8, respectively, at three representative operation frequencies: 400 MHz, 800 MHz and 1.6 GHz with level holding time of 2.5 ns, 1.25 ns and 0.625 ns, respectively. Although the discrete levels in the 8-bit staircase response of the optical links are indistinguishable due to noise effects, for photonic computing, NMSE is of more interest in quantifying the computing accuracy. In both Figs. 7 and 8, the traces of AWG (red), RF amplifier (pink), and MZM+PD (blue) follow the digital reference signal (green) closely at 400 MHz (first columns) and 800 MHz (second columns), while become struggled to follow at 1.6 GHz (third columns). At 1.6 GHz clock frequency with a holding time of 0.625 ns, the system response is unable to reach the targeted level before switching to the next symbol due to limited rise/fall time during the short level holding time. The highest deviation from the reference is observed when the input signal experiences a large transition between peak and valley, as shown in the third column of Figs. 7 and 8.



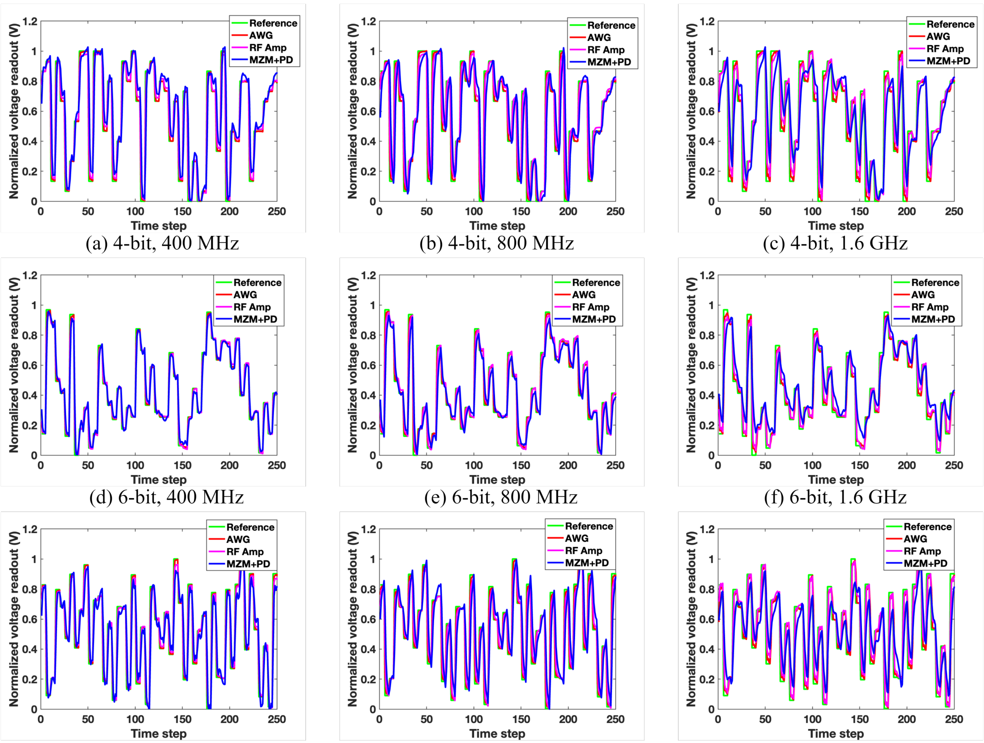
(a) 4-bit, 400 MHz
(b) 4-bit, 800 MHz
(c) 4-bit, 1.6 GHz
(d) 6-bit, 400 MHz
(e) 6-bit, 800 MHz
(f) 6-bit, 1.6 GHz
(g) 8-bit, 400 MHz
(h) 8-bit, 800 MHz
(i) 8-bit, 1.6 GHz

Fig. 7. MZM A test results at different clock frequencies and bit resolutions. Red curves represent the AWG signals. Pink curves represent the detected signals after the RF amplifier as the input signals to MZMs. Blue curves are the MZM+PD response containing the accumulated errors from AWG and RF amplifier

For optical computing tasks, since the deviation of the real-time transient response from the input signal is of most interest, we use the normalized mean square error (NMSE) as defined in Eq. (3) to quantify the accuracy of our EO MZMs, given by

$$NMSE = \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i {y_i}^2} \tag{3}$$

where $\hat{y}_i$ and y_i are the measured output response signal and targeted encoding signal, respectively. The calculated NMSE for two optical links, as well as AWG and RF amplifier responses at the testing clock frequencies from 100 MHz to 3.2 GHz, are illustrated in Fig. 9. The NMSE of AWG signal (blue) monotonically increases with the clock frequency from $\sim 10^{-6}$ to $\sim 10^{-3}$ due to the impact of system response time. For the RF amplifier and both MZMs responses, the NMSE shows a slightly decreasing trend with clock frequency at first while increasing again as frequency goes up to 1.6 GHz and 3.2 GHz. The slightly higher NMSE at 100 MHz is likely due to the low cut-off frequency of the system, mainly determined by the RF amplifier. At optimal



(a) 4-bit, 400 MHz (b) 4-bit, 800 MHz (c) 4-bit, 1.6 GHz

(d) 6-bit, 400 MHz (e) 6-bit, 800 MHz (f) 6-bit, 1.6 GHz

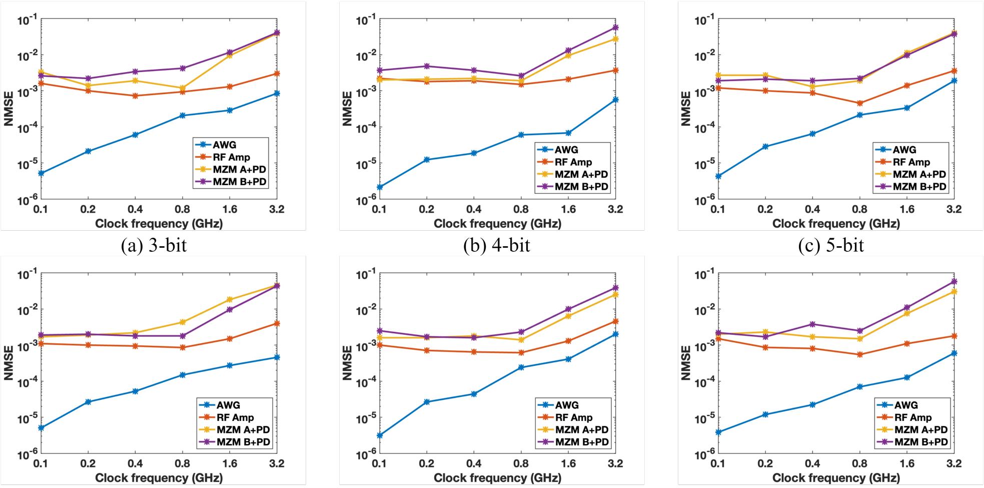
(g) 8-bit, 400 MHz (h) 8-bit, 800 MHz (i) 8-bit, 1.6 GHz

Fig. 8. MZM B test results at different clock frequencies and bit resolutions. Red, pink, and blue curves follow the same definition as MZM A testing results.

operating clock frequencies of 200 MHz, 400 MHz, and 800 MHz, where overall low NMSE are obtained, the RF amplifier already shows ~1 order of magnitude higher NMSE than AWG, meaning that the actual electrical signals that drive the MZMs are already deviated due to noise and nonlinearity of RF amplifier. The NMSE curves of MZM A+PD (yellow) and MZM B+PD (purple) are above the curve of the RF amplifier (orange) as a result of reduced SNR after the signals pass through MZM and PD. In the 1.6 GHz and 3.2 GHz cases, the level holding time of 0.625 ns and 0.3125 ns are shorter than the rise/fall time of both optical links while still longer than that of the RF amplifier. Therefore, the NMSE of both MZMs+PD shows a faster increase than the RF amplifier at those frequencies. For the entire optical link, the response time is the main limiting factor of the system performance and the major contribution to the NMSE at high clock frequency cases.

4.5. Power consumption

CV data was taken in order to estimate the energy consumption per switch, E_{sw}. The capacitance for MZMs A and B is measured to be 224 fF and 223 fF, giving rise to E_{sw}= 700 fJ/switch and 696.9 fJ/switch, respectively between lowest and highest energy levels (worst case scenario). Although MZM B has a shorter PS, it exhibits comparable capacitance due to higher doping concentrations of the PN junction. In a digital communication channel, for non-return-to-zero (NRZ) modulation format, the energy per bit can be calculated by $E_{bit} = \frac{1}{4}CV_{pp}{}^2$ [23] considering equal possibility for 0-0, 0-1, 1-0, 1-1 transitions, where "0" and "1" represents low energy level and high energy level, respectively. The PAM-N signal has N^2 possible transitions and $\log_2 N$ bits per symbol, with a voltage difference of $\frac{1}{N-1}V_{pp}$ between neighboring levels.



(a) 3-bit (b) 4-bit (c) 5-bit

(d) 6-bit (e) 7-bit (f) 8-bit

Fig. 9. NMSE results at different clock frequencies for 3~8-bit resolution. NMSE of RF amplifier contains the contributions from AWG and RF amplifier. NMSE of MZM+PD contains the contributions from AWG, RF amplifier, and MZM+PD.

For a computing task at M-bit vertical resolution, it is equivalent to having PAM-N input signals with encoded bit levels evenly distributed within a voltage range of V_{pp}. For an arbitrary transition from the i^{th} voltage level to the j^{th} voltage level, the power consumption caused by the junction capacitor charging/discharging can be written as:

$$E_{i,j} = \frac{1}{2} C {V_{pp}}^2 (\frac{j-i}{N-1})^2, \tag{4}$$

where $i, j = 1, 2, \ldots, N$. The average energy consumption per switch E_{sw} with randomly distributed M-bit signals under Vpp can be estimated from averaging the power over all the N^2 possible transitions $E_{i,j}$, given by [24]:

$$E_{sw} = \frac{1}{N^2} C {V_{pp}}^2 \sum_{i=1}^{N-1} (N-1)(\frac{i}{N-1})^2 \tag{5}$$

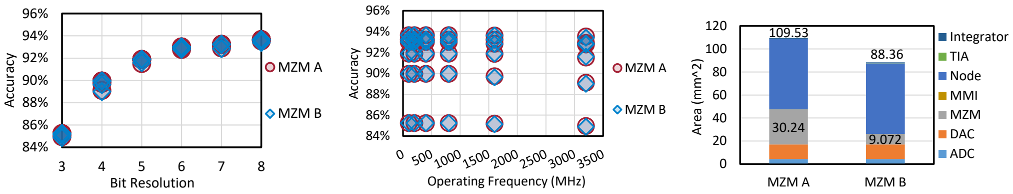
Based on the measured CV data, the energy consumption per bit and power consumption at 800 MHz clock frequency for 4, 6, and 8 bits are summarized in Table. 3. MZM A and B show similar energy per bit and power consumption due to similar capacitance values. For both MZMs tested, the averaged energy per switch E_{sw} and power consumption are lowest at 4-bit resolution and highest at 8-bit resolution because the chance of having large-range switches decreases as the number of bits increases.

Table 3. Average energy per switch and power consumption at 800MHz clock rate (symbol rate) for 4-, 6- and 8-bit resolution

Bit resolution (M)	4	6	8
Number of levels (N)	16	64	256
Clock rate (symbol rate)	800 MHz	800 MHz	800 MHz
Energy per switch (MZM A)	132.2 fJ/switch	120.4 fJ/switch	117.6 fJ/switch
Power consumption (MZM A)	105.8 μW	96.3 μW	94.0 μW
Energy per switch (MZM B)	131.6 fJ/switch	119.8 fJ/switch	117.1 fJ/switch
Power consumption (MZM B)	105.3 μW	95.9 μW	93.6 μW

5. Computing task validation

The experimentally characterized optical link, consisting of SL-MZM and an on-chip PD, provides essential terminal characteristics for the estimation of the performance of the tensor core in matrix multiplication. Built on a previously proposed tensor core architecture TeMPO [14], we model a real-time edge machine learning task running on TeMPO with 6 × 6 cores, each photonic tensor core size being 16 × 16. The results are shown in Fig. 10(a) and 10(b). This voice keyword spotting task [25] utilizes the Google Speech Commands dataset [26] for convolutional neural network (CNN) computing. To accommodate the MZM bit resolutions, a learned step size quantization-aware-training [27] approach is applied to the model, utilizing 3- to 8-bit symmetric signed per-tensor quantization for both weights and activations.



(a) (b) (c)

Fig. 10. Performance analysis of MZM A and B in real-time edge application CNN model and architecture simulation. (a) Impact of bit resolution on accuracy with noise. (b) Effect of operating frequency on accuracy with noise. (c) Architecture area simulation of both MZMs.

A +9% accuracy improvement is present when bit resolution is increased from 3-bit to 6-bit. Meanwhile, the accuracy improvement from 6- to 7-bit is negligible, meaning that the pursuit of higher bits may not be worth the power consumption increase needed to realize their operation. Operating frequencies have no significant impact on accuracy.

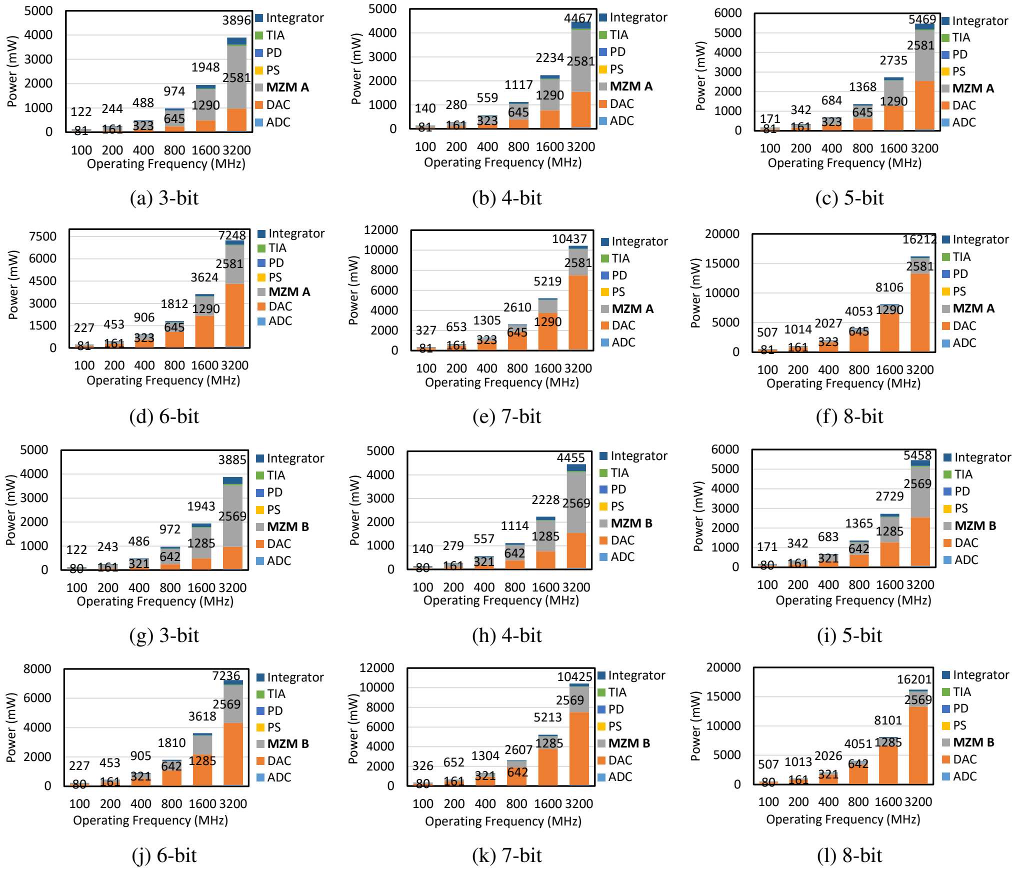
We also simulated the MZM power consumption and area cost using our previous work TeMPO [14] architecture setting. All the settings remain the same except for updating the MZM parameters to represent MZM A and B. The area and power breakdown for the architecture are shown in Fig. 10(c), and Fig. 11(a) to 11(l). In our simulation, the power consumption of a b_{in}-bit DAC is estimated as

$$P_{DAC}(b_{in},f) = P_{0,DAC} \cdot \frac{2^{b_{in}}}{b_{in}+1} \cdot f \tag{6}$$

and the power consumption of a b_{out}-bit ADC.

$$P_{ADC}(b_{out},f) = P_{0,ADC} \cdot b_{out} \cdot f \tag{7}$$

Here, $P_{0,DAC}$ and P_{ADC} are the reported power values of the DAC and ADC operating at their designed sampling rates and precision. DAC power scales linearly with the sampling frequency f and exponentially with the resolution b_{in}, while ADC power scales linearly with both the sampling frequency and the output bit resolution b_{out}. Therefore, lower bit resolution significantly contributes to power savings. The architecture's power consumption increases by an average of 1.4× for each increment in bit precision. When selecting the operating frequency, we prioritize higher frequencies, as they do not compromise model accuracy while significantly enhancing computational efficiency, measured in tera operations per second (TOPS). Under the same architecture size, a frequency of 3.2 GHz can result in a doubling of TOPS compared to 1.6 GHz (From 29.49 TOPS to 58.98 TOPS). Therefore, after considering all factors, MZM B operating at 3.2 GHz with 6-bit resolution is selected as the optimal choice.



(a) 3-bit (b) 4-bit (c) 5-bit

(d) 6-bit (e) 7-bit (f) 8-bit

(g) 3-bit (h) 4-bit (i) 5-bit

(j) 6-bit (k) 7-bit (l) 8-bit

Fig. 11. Power analysis of MZM A and B in architecture simulation using TeMPO [14] (a) - (f) **MZM A** architecture power simulation results at various operating frequencies and bit resolutions. (g) - (l) **MZM B** architecture power simulation results at different operating frequencies and bit resolutions.

With this architectural configuration, we achieve 58.98 TOPS, a power consumption of 7.32 W, an area of 88.36 mm^2, 8.06 TOPS/W in energy efficiency, and 0.67 TOPS/mm^2 in area efficiency. Our results demonstrate a 20× improvement in energy efficiency (0.4 TOPS/W to 8.06 TOPS/W) and 3.35× improvement in area efficiency (0.2 TOPS/mm^2 to 0.67 TOPS/mm^2) compared to NVIDIA V100 GPU [28].

6. Conclusion

This work thoroughly discusses the proof-of-concept of two computing units, each comprising an RF amplifier, a slow-light Mach-Zehnder modulator (SL-MZM) with a short PS length, and a compact on-chip Ge PD. The linearity was characterized, yielding a 6-bit error-free performance for both SL-MZMs under the chosen operating conditions, with partial nonlinearity cancellation observed between the RF amplifier and the on-chip SL-MZMs. While noise is inherent in all electronic components, the noise from the RF amplifier was observed to be the most detrimental as it causes fluctuations in optical signal intensity within the optical computing path, suggesting an optimal electro-optical conversion efficiency of the MZM for the best signal-to-noise ratio (SNR) at the physical layer and the lowest NMSE for computing accuracy. The NMSE of the computing links was evaluated across 4 to 8 bits of precision, with clock frequencies varying from 200 MHz to 3.2 GHz. Notably, a low NMSE of 0.0305 was achieved with 8-bit precision at a 3.2 GHz clock frequency using MZM A. The on-chip MZM and PD together exhibited a finite rise and fall time of approximately 0.8 ns, indicating that device bandwidth limits the NMSE to a clock frequency of ∼1.6 GHz. At a clock frequency of 800 MHz, we demonstrated an exceptionally low NMSE of 0.0018 for the computing link using MZM B, with a PS length of 150 μm. These results suggest a promising pathway for scaling up on-chip photonic computing using compact SL-MZMs and PDs. The voice keyword spotting task using the Google Speech Commands dataset showed an increase in computing accuracy up to 6-bit operation. When TOPS and power consumption are considered, the optimal computing choice studied in this work is MZM B, with a 6-bit resolution operating at 3.2 GHz. Our work demonstrates a 20× improvement in energy efficiency and a 3.35× improvement in area efficiency compared to NVIDIA V100 GPU [28].

Funding. AIM Photonics (FA8650-21-2-1000).

Acknowledgments. A. Begović would like to thank AIM Photonics for support in funding and chip space on the Costanza MPW run. M. Zhang would like to thank the New York State Focus Center for support via funding. N. Gangi would like to acknowledge the support of the DoD SMART Scholarship program.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. N. Peserico, X. Ma, B. J. Shastri, *et al.*, "Photonic tensor core for machine learning: a review," in *Emerging Topics in Artificial Intelligence (ETAI) 2022*, vol. 12204 G. Volpe, J. B. Pereira, D. Brunner, *et al.*, eds., International Society for Optics and Photonics (SPIE, 2022), p. 1220407.
2. M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," Appl. Phys. Rev. **7**(3), 031404 (2020).
3. X. Ma, R. L. T. Schwartz, B. Jahannia, *et al.*, "Fully integrated photonic tensor core for neural network applications," in *2023 IEEE Photonics Conference (IPC)*, (2023), pp. 1–2.
4. K. Amsalu and S. Palani, "A review on photonics and its applications," Mater. Today: Proc. **33**, 3372–3377 (2020).
5. Y. Jiang, W. Zhang, X. Liu, *et al.*, "Photonic micro-ring tensor core for parallel and shared batch processing," in *2023 Optical Fiber Communications Conference and Exhibition (OFC)*, (2023), pp. 1–3.
6. J. Feldmann, N. Youngblood, and M. Karpov, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**(7840), 52–58 (2021).
7. X. Fang and L. Yang, "Thermal effect analysis of silicon microring optical switch for on-chip interconnect," J. Semicond. **38**(10), 104004 (2017).
8. Z. Wang, D. Ming, and Y. Wang, "Resolving the scalability challenge of wavelength locking for multiple micro-rings via pipelined time-division-multiplexing control," Opt. Express **30**(14), 24984–24994 (2022).
9. O. Jafari, W. Shi, and S. Larochelle, "Mach-zehnder silicon photonic modulator assisted by phase-shifted bragg gratings," IEEE Photonics Technol. Lett. **32**(8), 445–448 (2020).
10. T. Baba, H. C. Nguyen, and N. Yazawa, "Slow-light Mach-Zehnder modulators based on si photonic crystals," Sci. Technol. Adv. Mater. **15**(2), 024602 (2014).
11. S. R. Anderson, A. Begovic, H. Jiang, *et al.*, "Compact slow-light integrated silicon electro-optic modulators with low driving voltage," IEEE Photonics Technol. Lett. **35**(13), 697–700 (2023).
12. O. Jafari, W. Shi, and S. LaRochelle, "Efficiency-speed tradeoff in slow-light silicon photonic modulators," IEEE J. Sel. Top. Quantum Electron. **27**(3), 1–11 (2021).

13. C. Han, M. Jin, and Y. Tao, "Recent progress in silicon-based slow-light electro-optic modulators," Micromachines **13**(3), 400 (2022).
14. M. Zhang, D. Yin, and N. Gangi, "TeMPO: Efficient time-multiplexed dynamic photonic tensor core for edge AI with compact slow-light electro-optic modulator," J. Appl. Phys. **135**(22), 223105 (2024).
15. Y. Yuan, "A 100 GB/s PAM4 two-segment silicon microring resonator modulator using a standard foundry process," ACS Photonics **9**(4), 1165–1171 (2022).
16. N. M. Fahrenkopf, "The aim photonics mpw: A highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits," IEEE J. Sel. Top. Quantum Electron. **25**(5), 1–6 (2019).
17. B. A. Marquez, J. Singh, and H. Morison, "Fully-integrated photonic tensor core for image convolutions," Nanotechnology **34**(39), 395201 (2023).
18. E. Dulkeith, F. Xia, and L. Schares, "Group index and group velocity dispersion in silicon-on-insulator photonic wires," Opt. Express **14**(9), 3853–3863 (2006).
19. J. Witzens, "High-speed silicon photonics modulators," Proc. IEEE **106**(12), 2158–2182 (2018).
20. *p-n Junctions*, (John Wiley & Sons, Ltd, 2006), pp. 77–133.
21. Y. Yuan, S. Cheung, T. Van Vaerenbergh, *et al.*, "A 7-bit precision linearized mach-zehnder interferometer for high accuracy optical neural networks," in *2023 Opto-Electronics and Communications Conference (OECC)*, (IEEE, 2023), pp. 1–3.
22. Z. Zhou and G. S. La Rue, "A 12-bit nonlinear dac for direct digital frequency synthesis," IEEE Trans. Circuits Syst. I **55**(9), 2459–2468 (2008).
23. R. G. Jesuwanth Sugesh and A. Sivasubramanian, "Modelling and analysis of a corrugated PN junction phase shifter in silicon MZM," Silicon **14**(6), 2669–2677 (2022).
24. R. Dubé-Demers, S. LaRochelle, and W. Shi, "Ultrafast pulse-amplitude modulation with a femtojoule silicon photonic modulator," Optica **3**(6), 622–627 (2016).
25. R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," (2018).
26. P. Warden, "Google speech commands dataset," (2017). Google Research Blog, accessed August 20, 2024.
27. S. K. Esser, J. L. McKinstry, D. Bablani, *et al.*, "Learned step size quantization," (2020).
28. J. Choquette, O. Giroux, and D. Foley, "Volta: Performance and programmability," IEEE Micro **38**(2), 42–52 (2018).