A 7-bit Precision Linearized Mach-Zehnder Interferometer for High Accuracy Optical Neural Networks

Yuan Yuan, Stanley Cheung, Thomas Van Vaerenbergh, Yiwei Peng, Yingtao Hu, Geza Kurczveil, Zhihong Huang, Di Liang, Wayne V. Sorin, Marco Fiorentino, and Raymond G. Beausoleil
Hewlett Packard Labs
Hewlett Packard Enterprise
Milpitas, CA 95035, USA
yuan.yuan@hpe.com

***Abstract*—A full-scale linearized MZI is proposed and experimentally validated by introducing an extremely overcoupled microring. It exhibits triangular-like transmission with a 3-bit improvement in precision while maintaining the same small number of phase shifters. The linearized MZI-based ONN demonstrates higher training accuracy.**

Keywords—bit precision, Mach-Zehnder interference, optical neural networks

I. Introduction

Artificial neural networks (ANNs) are biologically inspired algorithms that have emerged as a promising solution for intelligent computing to deliver human brain-like cognition. Many electronic neuromorphic systems, such as Intel's Loihi, IBM's TrueNorth, Google's TPU, Nvidia's Volta, and Qualcomm's Zeroth, have been demonstrated to provide specific hardware platforms. However, even with interconnected artificial neurons tailored for ANNs, these electronic systems still face intrinsic limitations in bandwidth and energy efficiency. On the other hand, optical interference allows efficient matrix-vector multiplication passively at the speed of light. Therefore, in theory, optical neuromorphic systems are superior to state-of-the-art neuromorphic electronics, which can operate several orders of magnitude faster (> 50 Gb/s) with lower power consumption. Lately, there have been numerous works on high-performance optical neural networks (ONNs), where an essential requirement is the ability to perform linear regression [1,2]. Mach-Zehnder interferometers (MZI), the most common optical interference unit, have been widely used as a basic cell in ONNs [1]. By changing the phase difference between the two arms, the MZI exhibits a sinusoidal transmission function. For ONN weight encoding, such sinusoidal function requires a complex and expensive specially designed electric circuit, for example, a nonlinear digital-to-analog converter (DAC) [3], to compensate for nonlinearity. Otherwise, a feedback loop or a pre-calibrated lookup table is essential for each MZI, which will significantly increase the system complexity, latency, and energy consumption [4]. Moreover, since electronic control circuits do not have infinite precision [5], the sinusoidal response ultimately limits the bit precision of each MZI unit and further restricts the training correctness of the whole system. To alleviate this nonlinearity, a strongly overcoupled microring is introduced on one arm to compensate for the sublinear sinusoidal response via its superlinear phase variation. The overcoupled ring-assisted MZI (RAMZI) can be used as a linearized interference unit with higher bit precision.

II. Device Design and Results

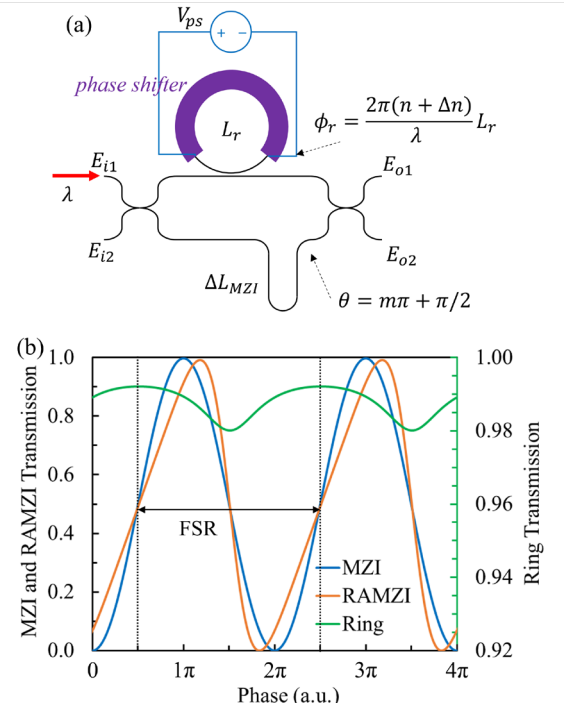

Fig. 1. (a) Schematic of the linearized RAMZI. (b) Calculated transmission versus phase change for MZI (blue), RAMZI (orange), and ring (green).

Previously, some MZI modulators have been reported to improve the linearity in the radio frequency (RF) domain, known as spurious free dynamic range (SFDR), by cancelling the third harmonic intensity distortion [6,7]. However, it is a small-signal solution that only works in a partial range. It is also unsuitable for scaling up MZI meshes due to the complicated modulation control of multiple phase shifters. In contrast, we propose a full-scale large-signal solution that covers the complete phase range from level "0" to "1". Fig. 1(a) shows a schematic of the linearized RAMZI, consisting of two 50%-50% couplers, one overcoupled all-pass ring, one waveguide delay arm, and one phase shifter locates on the ring. The roundtrip phase change of the ring $\phi_r = 2\pi(n + \Delta n)L_r/\lambda$, where n is the waveguide refractive index, Δn is the refractive index change caused by the phase shifter, L_r is

$$\begin{bmatrix} E_{o1} \\ E_{o2} \end{bmatrix} = \begin{bmatrix} \sqrt{0.5} & j\sqrt{0.5} \\ j\sqrt{0.5} & \sqrt{0.5} \end{bmatrix} \begin{bmatrix} \exp(j(\pi+\phi_r)) \dfrac{a - t\exp(-j\phi_r)}{1 - at\exp(j\phi_r)} & 0 \\ 0 & \exp(j\theta) \end{bmatrix} \begin{bmatrix} \sqrt{0.5} & j\sqrt{0.5} \\ j\sqrt{0.5} & \sqrt{0.5} \end{bmatrix} \begin{bmatrix} E_{i1} \\ E_{i2} \end{bmatrix} \quad (1)$$

the perimeter of the ring, and λ is the optical wavelength. The MZI delay arm ΔL_{MZI} introduces a quarter phase delay $\theta = m\pi + \pi/2$, where m is integer. Therefore, the resonate and off-resonate points of the overcoupled ring coincides with the midpoints of the sinusoidal wave, as shown in Fig. 1(b). An overcoupled microring with optimal field coupling coefficient $\kappa = 0.92$ is applied to introduce suitable phase changes while guaranteeing very small transmission amplitude variations, shown as the green curve in Fig. 1(b). The output field of the RAMZI can be expressed as Eq. (1), where a is the roundtrip field transmission of the ring, and t is the field transmission coefficient of the bus-ring coupler. By optimizing the field coupling coefficient of the bus-ring coupler, κ, the RAMZI transmission can be linearized. For a lossless bus-ring coupler, $\kappa^2 + t^2 = 1$. The calculated optical power transmission versus phase change for conventional MZI (blue curve) and proposed RAMZI (orange curve) is shown in Fig. 1(b), where the ring parameters $a = 0.99$ and $\kappa = 0.92$. The RAMZI exhibits a triangular-like response. To quantify the linearity, the linear regression of MZI (blue dash line) and RAMZI (orange dash line) from transmission "0" to "1" are plotted in Fig. 2(a). The standard deviation (σ) of MZI is ~ 0.043, while the σ of overcoupled RAMZI is improved to ~ 0.007, which is more than 6 times smaller. The associated residuals of the linear regression are plotted in Fig. 2(b). The 4-, 5-, 6-, and 7-bit least significant bit (LSB) intervals are also plotted to aid in the intuitive assessment of the residuals, the conventional MZI enables a 4-bit precision, whereas the RMAZI offers a 7-bit precision.

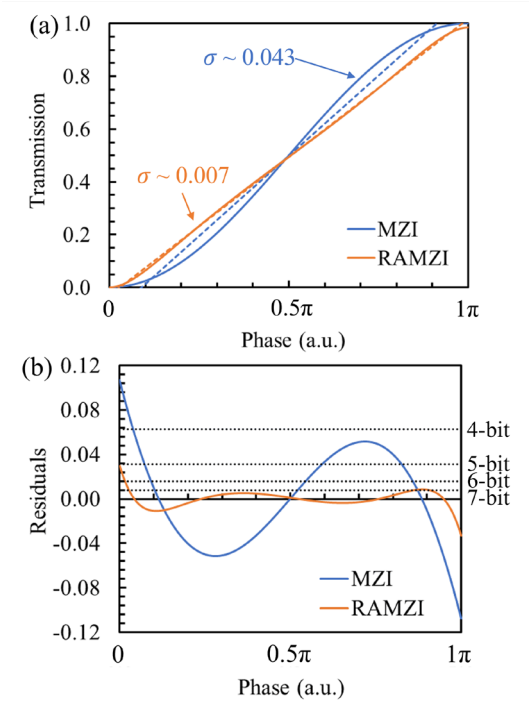

Fig. 2. (a) Linear regression from transmission "0" to "1", and (b) the associated linear regression residuals for MZI and RAMZI.

The linearized RAMZI design has been experimentally validated by a demonstrated device, whose microscope image is shown in Fig. 3(a). The tested RAMZI is part of a (de)-interleaver design of experiments (DOE), which is based on a 300 nm-thick SOI [8]. The designed ring circumference L_r is 1200 μm and the arm delay length ΔL_{MZI} is 600 μm. By tuning the optical wavelength to 1308.8 nm, the delay phase from ΔL_{MZI} satisfies the quarter phase condition. Figure 3(b) shows the measured optical transmission versus ring heater power, which exhibits a triangular-like shape. The linear regression line from level "0" to level "1" is shown as an orange dashed line, indicating the good linearity of the tested RAMZI. The simulated transmission of the tested device is illustrated in Fig. 3(c), which agrees well with the measured results. Here, the field coupling coefficient, $\kappa \sim 0.985$, is higher than the optimal value, therefore the linearity of the RAMZI can be further improved with an optimal κ.

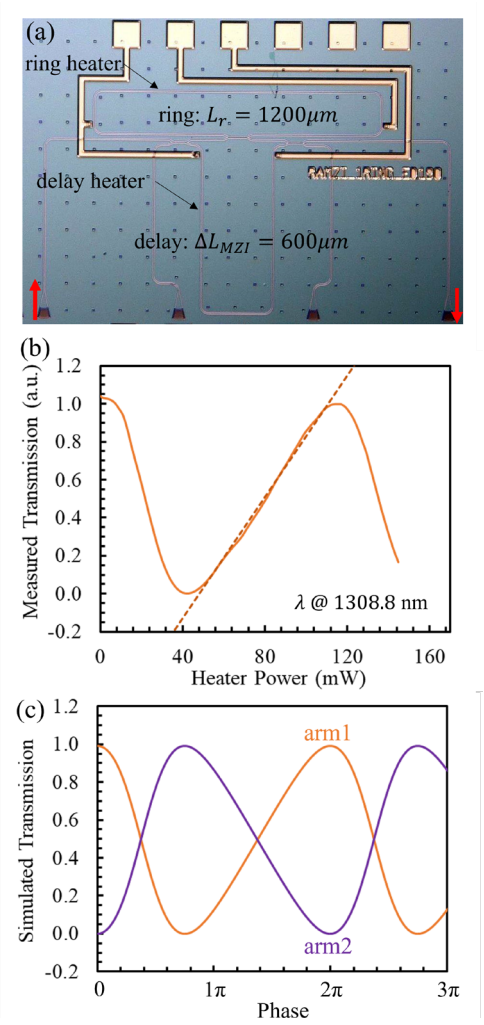

Fig. 3. (a) Microscope image of the tested RAMZI with heaters. (b) Measured optical transmission versus ring heater power at 1308.8 nm. (c) Simulated optical transmission versus phase change.

III. Discussion and Conclusion

This work demonstrates a linearized RAMZI operating over the full-scale range. Relative to the sinusoidal response, the triangular transmission provides a more than 6-fold reduction in residual error, resulting in a 3-bit improvement in accuracy. 7-bit precision can be easily achieved without any nonlinear compensation control circuit, which is sufficient for many neural network models [9]. Moreover, it does not add any extra phase tuning elements, only one phase shifter on the overcoupled ring, making it suitable for large-scale integration like Reck and Clements MZI meshes. Such RAMZI can replace common MZIs in ONNs to further improve training correctness. To validate it, a chip-level ONNs simulation package *Neuroptica* [10] is utilized to classify a randomly generated ring-shape planar dataset shown in Fig. 4(a), where

the ring-shape red points are “label 0” and the other blue points are “label 1”. ONNs based on conventional MZIs and linearized RAMZIs were simulated with five Clements mesh layers and each layer has five input waveguides. The same on-chip training with in-situ backpropagation using the adjoint field method [11] and ADAM optimizer [12] was applied with an optimization step size of 0.02. The simulated cross-entropy loss ($\mathcal{L}$) of ONNs is shown in Fig. 4(b), the ONN using RAMZI achieves faster convergence and about 60% lower $\mathcal{L}$ compare to the ONN with MZI. The predicted decision planar boundary of MZI and RAMZI ONNs is shown in Fig. 4(c). The predicted boundary based on the linearized RAMZI is much closer to the labelled ring-shape classification.

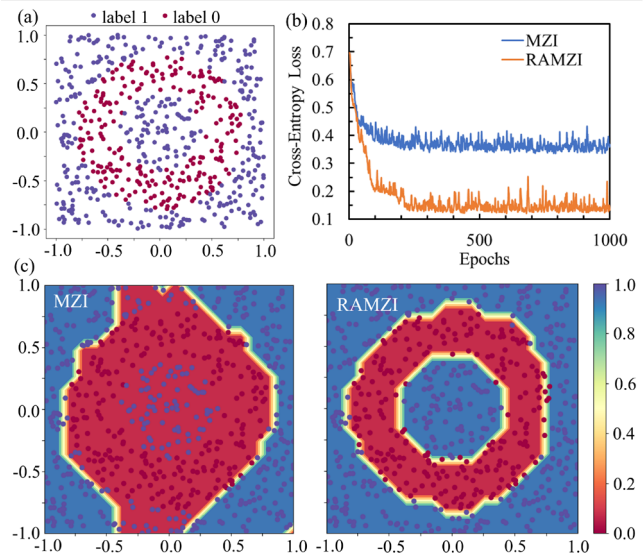

Fig. 4. Ring planar data classification using *Neuroptica* [10] with ADAM optimization step size of 0.02: (a) randomly generated ring-shape planar dataset, (b) cross-entropy loss versus epoch, and predicted decision planar boundary of (c) MZI and (d) RAMZI ONNs.

Consequently, the linearized RAMZI yields a 3-bit higher bit precision compared to the conventional MZI while maintaining the same scalability, which can significantly reduce the complexity and energy cost of ONN control circuits. The linearized RAMZI has been experimentally validated, and its 7-bit precision offers sufficient resolution for many neural network models. In addition, the RAMZI-based ONN exhibits faster convergence and higher training accuracy under the same optimization step size.

REFERENCES

[1] Y. Shen, et al. “Deep learning with coherent nanophotonic circuits,” Nat. Photon 11(7), 441-446 (2017).

[2] B. J. Shastri, et al. “Photonics for artificial intelligence and neuromorphic computing,” Nat. Photon 15(2), 102-114 (2021).

[3] Z. Zhou, et al. “A 12-bit nonlinear DAC for direct digital frequency synthesis,” IEEE Trans. Circuits and Syst. I: Regul. Pap. 55(9), 2459-2468 (2008).

[4] W. Zhang, et al. “Silicon microring synapses enable photonic deep learning beyond 9-bit precision,” Optica 9(5), 579-584 (2022).

[5] Y. Ehrlichman, et al. “Improved digital-to-analog conversion using multi-electrode Mach–Zehnder interferometer,” J. Light. Technol. 26(11), 3567-3575 (2008).

[6] X. Xie, et al. “Linearized mach-zehnder intensity modulator,” IEEE Photon. Technol. Lett. 15(4), 531-533 (2003).

[7] J. Cardenas, et al. “Linearized silicon modulator based on a ring assisted Mach Zehnder inteferometer,” Opt. Express 21(19), 22549-22557 (2013).

[8] S. Cheung, et al. “Ultra-power-efficient heterogeneous III–V/Si MOSCAP (de-) interleavers for DWDM optical links,” Photonics Res. 10(2), A22-A34 (2022).

[9] B. Jacob, et al. “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in Proceedings of the IEEE conference on computer vision and pattern recognition (2018) pp. 2704–2713.

[10] B. Bartlett. “Nanophotonic Neural Network Simulator,” https://github.com/fancompute/neuroptica.git.

[11] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic neural networks through in situ backpropagation and gradient measurement,” Optica 5, 864–871 (2018).

[12] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980 (2014).