Towards a High-density Photonic Tensor Core Enabled by Intensity-modulated Microrings and Photonic Wire Bonding: supplemental document

This document provides supplementary information to "Towards a High-density Photonic Tensor Core Enabled by Intensity-modulated Microrings and Photonic Wire Bonding". Details are provided about design, fabrication, simulations and experimental procedures for the devices presented in the main article.

1. DESIGN OF THE PROPOSED MODULATOR

The design of the proposed intensity-modulation-based microring modulator (IM-MRM) follows standard design rules of the microring resonator (MRR) described in Ref. [1], except that a two-point Mach-Zehnder interferometer-based (MZI) coupler is utilized for the ring-bus coupling. Transfer Matrix Method (TMM) is used to calculate the transmission and reflection between the MZI and MRR, as shown in Figure S1(a). Equations S1 and S2 shown below present the transfer matrix and transfer function between the input (E_{in1} and E_{in2}) and output (E_{thru1} and E_{thru2}) electric fields of the MZI [2]:

$$\begin{bmatrix} E_{\text{thru1}} \\ E_{\text{thru2}} \end{bmatrix} = \begin{bmatrix} T_1 & K_2 \\ K_1 & T_2 \end{bmatrix} \begin{bmatrix} E_{\text{in1}} \\ E_{\text{in2}} \end{bmatrix}, \tag{S1}$$

$$E_{\text{in2}} = \alpha_{\text{MRR}} \exp(-j\delta) E_{\text{thru2}}, \tag{S2}$$

where α_{MRR} represents the single-pass amplitude transmission and $\delta = \beta L$ represents the single-pass phase shift in the MRR, β is the propagation constant and $L = 2\pi R$, and R is the radius of the MRR. T_{i} and K_{i} represent the transmission from the i$^{\text{th}}$ input port of the MZI to the through-port and cross-port output (i = 1, 2), respectively, and can be calculated using the following equations:

$$T_1 = t_1 t_2 \alpha_{\text{arm1}} \exp(-j(\phi_1 + \Delta\phi_1)) + k_1 k_2 \alpha_{\text{arm2}} \exp(-j\phi_2), \tag{S3}$$

$$T_2 = t_1 t_2 \alpha_{\text{arm2}} \exp(-j\phi_2) + k_1 k_2 \alpha_{\text{arm1}} \exp(-j(\phi_1 + \Delta\phi_1)), \tag{S4}$$

$$K_1 = k_1 t_2 \alpha_{\text{arm2}} \exp(-j\phi_2) + t_1 k_2 \alpha_{\text{arm1}} \exp(-j(\phi_1 + \Delta\phi_1)), \tag{S5}$$

$$K_2 = k_1 t_2 \alpha_{\text{arm1}} \exp(-j(\phi_1 + \Delta\phi_1)) + t_1 k_2 \alpha_{\text{arm2}} \exp(-j\phi_2), \tag{S6}$$

where t_{i} and k_{i} represent the transfer functions of the electric field coupled to the through-port and cross-port of the i$^{\text{th}}$ coupler (i = 1, 2), respectively, and ϕ_1 and ϕ_2 are the phase responses of the upper and lower MZI arms. $\Delta\phi_1$ represents the phase change induced by manipulating the index modulation element in the upper arm of the MZI coupler, and α_{arm1} and α_{arm2} are the amplitude transmissions, accordingly. The electric field transfer function at the through port of the device is [2]:

$$\frac{E_{\text{thru1}}}{E_{\text{in1}}} = \frac{T_1 - (T_1 T_2 - K_1 K_2)\alpha_{\text{MRR}} \exp(-j(\delta + \Delta\phi_2))}{1 - T_2 \alpha_{\text{MRR}} \exp(-j(\delta + \Delta\phi_2))}. \tag{S7}$$

If we assume a loss-less MZI coupler in the design:

$$T_1 T_2 - K_1 K_2 = \exp(-j(\phi_1 + \phi_2 + \Delta\phi_1)), \tag{S8}$$

$$T_1 = \sin((\phi_1 - \phi_2 + \Delta\phi_1)/2)\exp(-j(\phi_1 + \phi_2 + \Delta\phi_1 + \pi)/2), \tag{S9}$$

$$T_2 = \sin((\phi_1 - \phi_2 + \Delta\phi_1)/2)\exp(-j(\phi_1 + \phi_2 + \Delta\phi_1 - \pi)/2). \tag{S10}$$

Therefore, according to Equations S7 and S8, the resonance condition (when $\phi_1 + \phi_2 + \Delta\phi_1 + \delta + \Delta\phi_2 = 2\pi m$, where m is integer) of the device depends on the variation of both phase changes $\Delta\phi_1$ and $\Delta\phi_2$. While, the transmission amplitude, according to Equations S9 and S10, depends on the additional phase change ($\Delta\phi_1$) between MZI's two arms. The interference modifies the effective coupling ratio in the MZI coupler, thus changing the extinction ratio (ER) at the resonance peak.

As shown in Figure S1(a), in our proposed design, an unbalanced MZI coupler is used for input/output coupling. Therefore, the coupling coefficient is wavelength-dependent. When a broadband phase shift $\Delta\phi_1$ on the upper arm of the MZI is actuated, the sinusoidal spectral response of the MZI coupler shifts in wavelength, thereby changing the effective coupling ratio to the MRR [3]. At the same time, the resonance condition will change accordingly. Therefore, the index modulation element in the MRR serves as the wavelength drift compensator to offset the drift of the resonance peak with a negligible impact on the transmission amplitude.

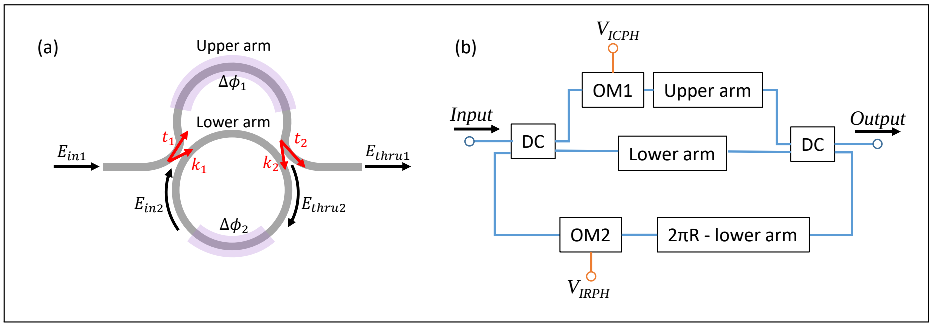

Fig. S1. (a) Schematic of the proposed IM-MRM. Purple areas represent index modulation elements, providing $\Delta\phi_1$ and $\Delta\phi_2$ phase shifts to the MZI coupler and the MRR, respectively. (b) Proposed IM-MRM custom compact model in INTERCONNECT. Two DC blocks are used to couple the light on and off the MRR, two OM elements are used for the index modulation, and 3 passive waveguide elements are used to build the MZI and MRR. Blue lines represent optical interconnects and orange lines represent electrical interconnects.

In our proposed design, as shown in Figure S1(b) of the paper, the index modulation element in the MRR has been placed in the lower arm of the MZI to balance the refractive indices between two arms and to reduce the footprint by sharing the ground. This design realizes the intensity modulation at a fixed wavelength, but thermal crosstalk between upper and lower arms is observed. Moreover, modulating the index modulation element in the lower arm not only changes the resonance condition, but changes the effective coupling ratio in the MZI coupler, thus changing the ER slightly. Luckily, it happens with a lower efficiency compared with the upper arm's modulation efficiency which has a longer length. In the future design, this will be optimized.

The simulation model for the proposed IM-MRM was developed using Lumerical Suite [4]. Lumerical MODE Solutions and DEVICE were first used to obtain the effective index change and loss as a function of bias voltage for the N-doped resistive heater. The data was then imported to INTERCONNECT to build the optical modulator (OM) element with parameterizable lengths and bias voltages. Figure S1(b) shows the custom compact model for the proposed all-pass IM-MRM. Two directional coupler (DC) blocks and three passive waveguide sections, as well as two OM elements, were used to build the MZI coupler and the MRR. It is worth noting that the DC and the passive waveguide are parameterizable. Therefore, the compact model can be adjusted with the coupling strength, the lengths of the MZI coupler arms, the radius of the MRR, and the fill-factor of the N-doped resistive heater to obtain the best performance. At last, we changed the applied voltages, V_{ICPH} and V_{IRPH}, independently and recorded the resonance wavelength drift accordingly to find the wavelength compensation voltage pair for locking the resonance peak of the model. By applying the obtained voltage pairs to V_{ICPH} and V_{IRPH} ports, the intensity modulation at a fixed wavelength is realized.

2. LINEAR PREDISTORTION

The optical measurement setup consists of a tunable off-chip laser source (Agilent 81682A) with power meters (Agilent 81635A) as the detector, an optical fiber array (PLC Connections, Columbus, OH) to couple the light on and off the chip, and a thermally-tuned stage controlled by a temperature controller (SRS LDC501). Electrical control of ICPH and IRPH were performed by a source meter (Keithley 2604B) in the constant voltage mode.

Since our proposed IM-MRM is operated leveraging the thermo-optic effect, the relationship between the applied voltage and the transmitted power is non-linear. Non-linear response of the detected transmitted powers is obtained and shown in Figure S2(a) by sweeping the voltage pairs with a fixed V_{ICPH} step. The power levels are collected at the minimum and maximum of the power range with 4-bit precision. To obtain a linear relationship between the applied voltage pair and the transmitted power, a predistortion step is introduced. As depicted in Figure S2(b), according to the obtained detected power, a polynomial interpolation method is introduced which adds more points along the non-linear curve. By re-distributing 16 points with a fixed transmitted power level (yellow dots in Figure S2(b)), the required V_{ICPH} values are obtained. By applying the re-distributed voltage pair, generated by the predistortion step, to the IM-MRM, sixteen linearly distributed power levels are observed in Figure 4(c) of the main paper. This is important for the implementation of linear encoding/decoding with multiple wavelength channels.

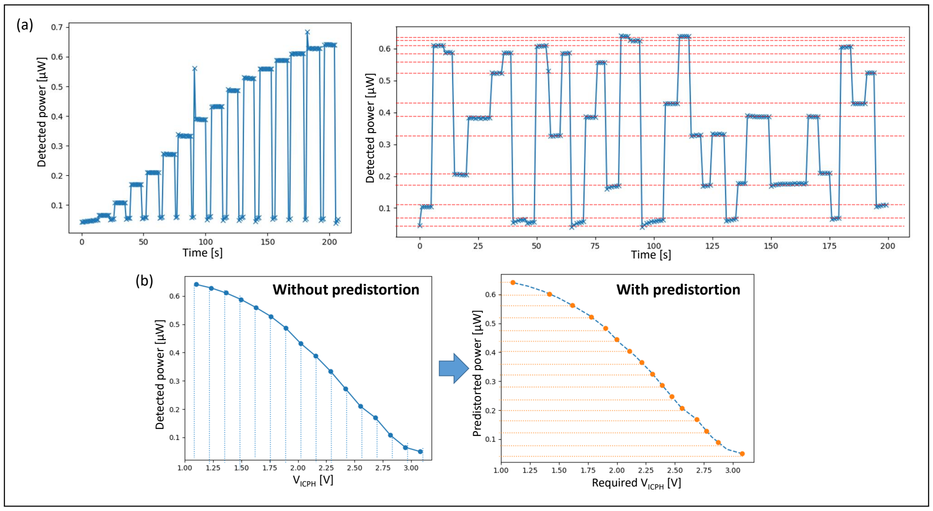

Fig. S2. (a) 16 power levels in consecutive ascending order of the all-pass IM-MRM. A non-linear relationship between the detected power level and the applied voltage pair is observed. (b) 16 power levels vs. applied voltages without and with the predistortion step. After predistortion, linearly distributed 16 power levels are observed.

3. ACHIEVABLE VALUES VS. WAVELENGTH CHANNEL SPACING

As described in Section 2.1 of the main paper, a custom compact model is developed for the proposed IM-MRM design using Lumerical INTERCONNECT to investigate the inter-channel crosstalk penalty. To visualize the power penalty vs. the wavelength channel spacing, we built two cascaded types (Type-I and Type-II) in INTERCONNECT and simulated with different wavelength channel spacing between the two resonance peaks. The wavelength channel spacing was modified by changing the radius of one of the IM-MRM models. Before the power penalty investigation, the intensity modulation capability of the individual IM-MRMs was tested by applying different voltage pairs. Simulated transmission response at the resonance peak was recorded accordingly and then used to build the relationship between the applied voltage pair and the transmitted power of each IM-MRM model. Figure refs3 presents simulated achievable value ranges for Channel-1 and Channel-2 with different wavelength channel spacing. Figure refs3(a) and refs3(b) represent the simulation results for Type-I and Type-II systems, respectively. For Type-I systems, the 3-dB penalty tolerable wavelength channel spacing is roughly $0.5\delta\omega$ due to its Lorentzian-shaped resonance peak resulted from the all-pass IM-MRM (see Figure 6(b) of the main paper). While for Type-II systems, thanks to the drop port transmission, the filtered power at the through port in Channel-1 can be detected at the drop port (see Figure 6(c) of the main paper), thus offering a smaller wavelength channel spacing ($\delta\omega = 0.2$) with 3-dB power penalty.

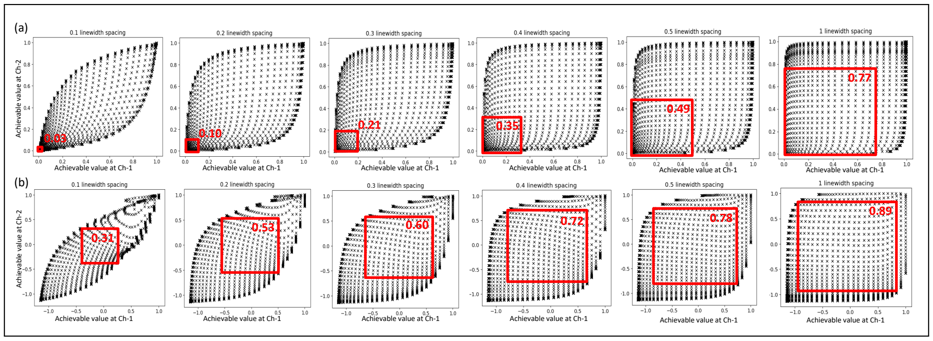

Fig. S3. Simulated achievable values for Channel-1 and Channel-2 of two different types of cascaded IM-MRM systems. (a) Type-I. (b) Type-II. The red box in the plot represents the usable range.

Similarly, achievable value range for wavelength-modulation-based MRM systems are investigated and presented in Figure S4. After replacing the IM-MRM model with the standard MRM model in INTERCONNECT, the transmitted power at Channel-1 and Channel-2 were recorded when wavelength channel spacing varies from $0.5\delta\omega$ to $4\delta\omega$. Only Type-II system is considered in the WM-MRM system. As can be seen, due to the wavelength drift, a larger wavelength channel spacing ($> 3\delta\omega$) is required for the 3-dB power penalty tolerance, which is in good agreement with simulation results in Ref. [5]. For $0.5\delta\omega$, there is no usable range due to the flat-top transmission at the drop port (first plot of Figure S4(a)) which is caused by the coherent interaction between the two resonance peaks.

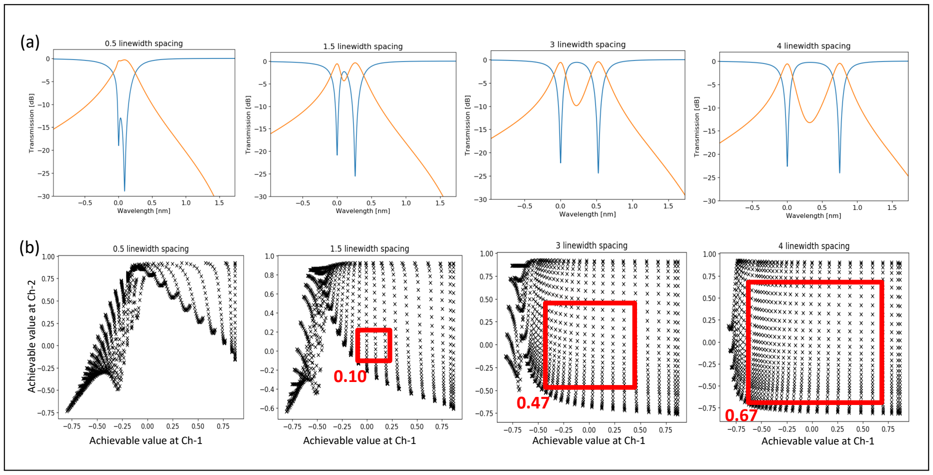

Fig. S4. (a) Simulated transmission spectra of Type-II WM-MRM systems with different wavelength channel spacing. Blue curves represent through-port transmissions and orange curves represent drop-port transmissions. (b) Simulated achievable values for Channel-1 and Channel-2 of WM-MRM systems with different wavelength channel spacing. The red box in the plot represents the usable range.

4. CO-SIMULATION PIPELINE AND HAND-WRITTEN DIGIT RECOGNITION RESULTS

To reduce the simulation time for hand-written digit recognition task, a co-simulation pipeline using Lumerical API is developed. The schematic of the co-simulation pipeline is presented in Figure S5. We use the aforementioned IM-MRM model in INTERCONNECT to build a convolution channel with WDM incoming signals to add the inter-channel crosstalk penalty between cascaded IM-MRM models. Transmission responses of IM-MRMs in modulation banks and weight banks are extracted for each model as a function of the voltage pairs applied for intensity modulation. Trained weights and biases (calculated using TensorFlow) for MNIST dataset are rounded to available values according to the transmission response of the weight banks obtained from INTERCONNECT and stored in Python. The rounding accuracy depends on the precision of the weight banks. For inference, new MNIST images are imported through the modulation banks, then weighted by the weight banks for matrix multiplication. The results are obtained by applying ReLU activation function and performing the remaining operations in average pooling and fully-connected layers. Compared with completing all the simulations in INTERCONNECT, using co-simulation pipeline can speed up the CNN simulation, especially for large-scale systems. Moreover, this pipeline can incorporate the inter-channel crosstalk penalty into the simulations, thus providing more realistic transmission responses for a cascaded system.

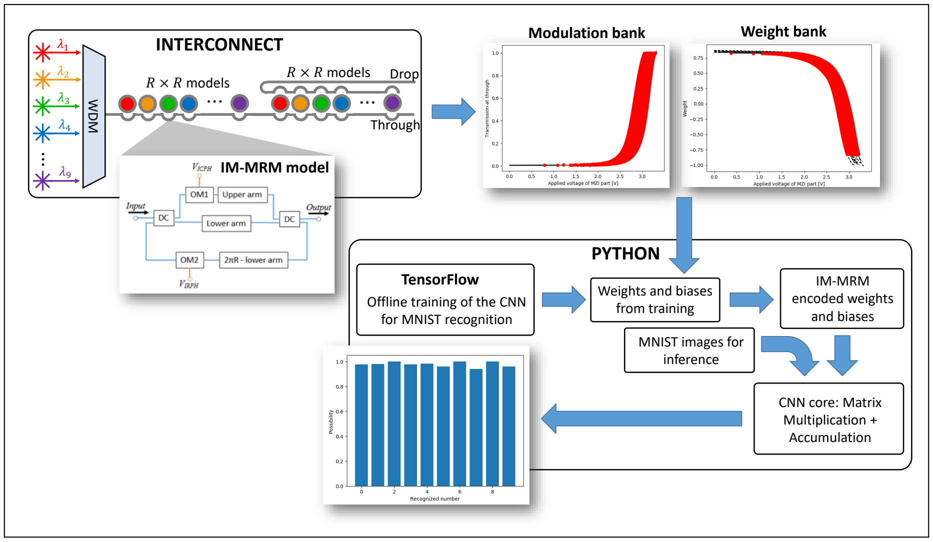

Fig. S5. Co-simulation pipeline used for hand-written digit recognition task.

The working procedure of the co-simulation pipeline is as follows. According to the dimensionality ($R \times R$) of the kernel in each channel (the number kernels is K), R^2 add-drop IM-MRM models are implemented and cascaded with varied radii in Lumerical INTERCONNECT via Python API. The radius of each model is adjusted depending on its operational wavelength from R^2 light sources. Transmission responses including the inter-channel crosstalk of the cascaded system are calculated in INTERCONNECT and then exported to Python. By applying different voltage pairs to individual IM-MRM models in the cascaded system, simulated power range is obtained by subtracting transmission responses from drop and through ports for each wavelength (as shown in Figure S6). A common power range is extracted according to the maximum achievable value for all models in the system (see Figure 7(c) of the paper). The common power range is then normalized and quantized based on the required precision and stored for encoding the kernel in Python. The relationship between each normalized power value and the applied voltage pair for individual IM-MRMs is saved as the mapping table which will be tracked when loading the target value to the system. Similarly, transmission responses of modulation banks are calculated for R^2 all-pass IM-MRM models and used for encoding input WDM signals. Convolution computations for the proposed CNN simulator are proceeded through matrix-matrix multiplication in Python. The subset of the input image is encoded as an input vector using the transmission data of the all-pass IM-MRM, and then multiplied by the kernel vector encoded using the weighting data of the add-drop IM-MRM. The single convolved result is generated by accumulating multiplication outputs and biases from K channels. By striding along the $H \times W$

input image with a stride of S, the image is transformed into a matrix of dimensionality of $KDR^2 \times [(H-R)/S+1][(W-R)/S+1]K$, where H and W are the height and width of input image including padding, and D is the number of input channels. Because only normalized input values smaller than 1 can be encoded using the modulation banks, the convolution result is first normalized and then goes through a ReLU activation function in Python. After the activation function, the image will pass through the second set of convolutional layer and activation (ReLU) function. A 2×2 average-pooling layer is utilized for invariance and direct down-sampling of the convolved features. Finally, a fully-connected layer is fed by flattening the pooled image, and the resultant vector is proceeded with the last fully-connected layer, where the result of the hand-written digit recognition task is obtained using SoftMax function. Convolutional parameters are summarized in Table S1.

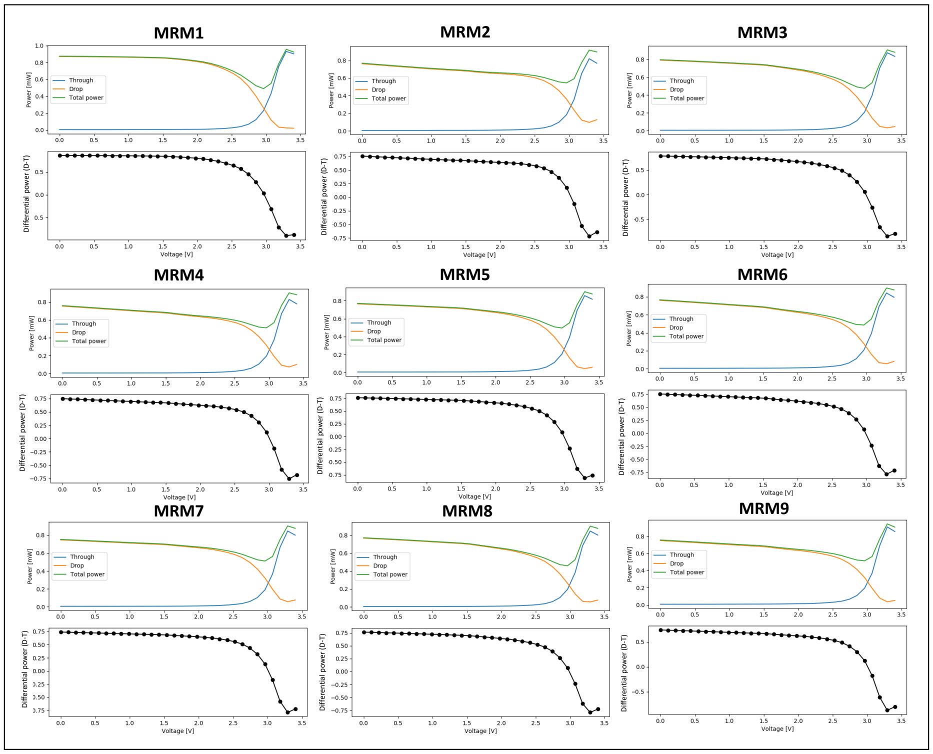

Fig. S6. Simulated transmitted power at through and drop ports, and differential power range vs. applied voltage pairs for the intensity modulation of 9 add-drop IM-MRM models in the cascaded system in INTERCONNECT. Blue curves represent the detected power at the through port, and orange curves represent the detected power at the drop port. Green curves show the sum of the detected power at both through and drop ports, indicating the IL of each IM-MRM. Black dots show the differential power calculated by subtracting the drop- and through–port powers for different applied voltage pairs.

Different precision of the transmission responses of the modulation and weight banks in the co-simulation pipeline (with 3-dB power penalty crosstalk), results in different overall accuracy. As shown in Figure S7, when the number of precision bits are changed from 3 to 10, the recognition accuracy increases. Considering the 8.5-bit precision recently achieved using IRPH-based MRMs, our proposed IM-MRM system shows an overall accuracy of ∼98%.

5. CO-PACKAGING OF THE PHOTONIC CHIP

We used a chip-on-board assembly method for co-packaging of the photonic chip. The three main steps for co-packaging are as follows. First, the photonic chip and V-grooves with single-mode fibres are directly mounted on a PCB substrate using UV curable epoxies (Norland Optical Adhesive NOA 86H). In the second step, photonic wire bonding is done using a photonic wire bonder (Vanguard Automation GmbH, SONATA1000). The third step is electrical wire bonding

Table S1. **Summary of convolutional parameters [6]**

Parameters	Meaning
H	Height of input images including padding
W	Width of input images including padding
D	Number of input channels
R	Kernel's edge length
K	Number of kernels
S	Stride

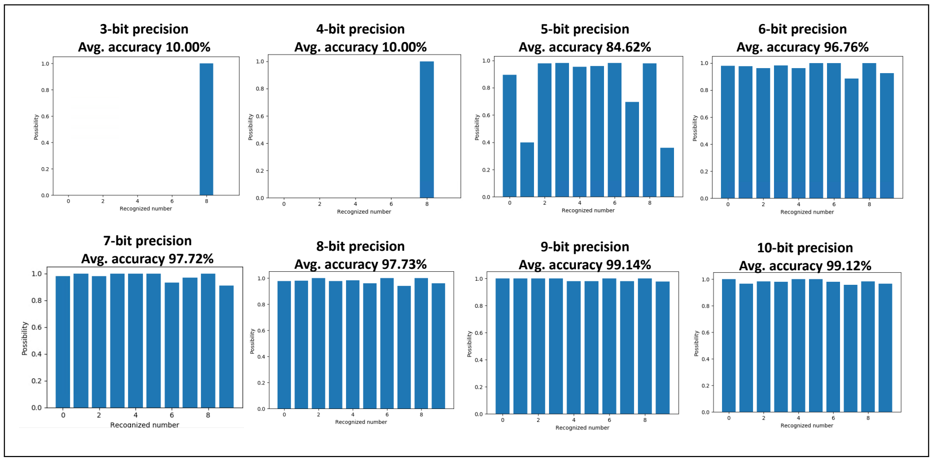

Fig. S7. Simulated accuracies of the hand-written digit recognition task for different bits of precision.

(EWB). At this step, the aluminum (Al) bond pads on the photonic chip are wire-bonded to the corresponding electroless nickel immersion gold pads on the PCB to implement chip-to-PCB electrical interconnects. An ultrasonic energy is used to attach an Al wire from the photonic chip pads to the PCB pads. We used the wedge-wedge bonder (Westbond 7476E) for electrical wire bonding.

Figure S8(a) shows the microscopic image of the photonic chip before the electrical wire bonding (EWB). The left-hand-side all-pass MRR (input MRR) serves as the input signal encoder, and the right-hand-side add-drop MRR filter (weight MRR) serves as the weight encoder, respectively. Figure S8(b) presents SEM images of the tip of the silicon taper with an oxide opening, which is used as the PWB interface to couple the light on and off the photonic chip. The chip was firstly sputtered with 5-nm-thick iridium (Leica EM ACE600) for SEM imaging. The roughness around the silicon taper indicates relatively high waveguide loss due to sidewall scattering. The trench around the waveguide taper is a side effect of the isotropic wet etching in the oxide opening process, which exposes the propagation mode to the sidewall due to a lower mode confinement. We believe the poor coupling efficiency should be attributed to these two factors, and better lithography processes may improve the performance.

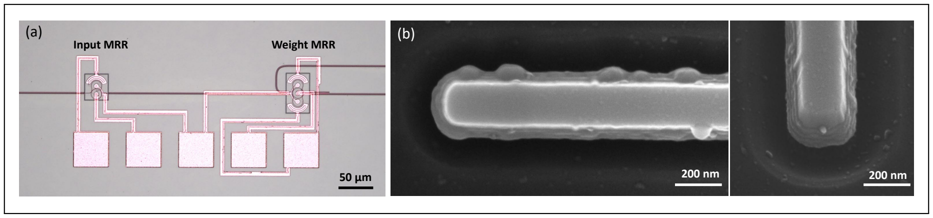

Fig. S8. (a) Microscopic image of the photonic chip before co-packaging. (b) SEM images of the tip of the silicon taper for the PWB interface.

REFERENCES

1. J. Heebner, R. Grover, and T. Ibrahim, *Optical microresonator theory* (Springer, 2008).
2. M. S. Hai, M. M. P. Fard, and O. Liboiron-Ladouceur, “A ring-based 25 gb/s dac-less pam-4 modulator,” IEEE J. Sel. Top. Quantum Electron. **22**, 123–130 (2016).
3. M. Popović, “Theory and design of high-index-contrast microphotonic circuits,” Ph.D. thesis, Massachusetts Institute of Technology (2008).
4. Ansys-Lumerical, “Lumerical,” https://www.lumerical.com/ (Accessed: 18 January 2022).
5. A. N. Tait, “Silicon photonic neural networks,” Ph.D. thesis, Princeton University (2018).
6. V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, “Digital electronics and analog photonics for convolutional neural networks (deap-cnns),” IEEE J. Sel. Top. Quantum Electron. **26**, 1–13 (2019).