AFRL-AFOSR-VA-TR-2024-0048

RANC: A Residue Arithmetic Nanophotonic Computer

El-Ghazawi, Tarek
THE GEORGE WASHINGTON UNIVERSITY
2121 I ST NW STE 601
WASHINGTON, DC,
US

11/30/2023
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20231130	Final	START DATE 20190801	END DATE 20221231

4. TITLE AND SUBTITLE
RANC: A Residue Arithmetic Nanophotonic Computer

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-19-1-0277	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Tarek El-Ghazawi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
THE GEORGE WASHINGTON UNIVERSITY
2121 I ST NW STE 601
WASHINGTON, DC
US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203	AFRL/AFOSR RTB1	AFRL-AFOSR-VA-TR-2024-0048

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Due to the end of Moore's law and Dennard scaling, feature reduction and higher speed of clocking are seizing to be the source for higher computer performance. Therefore, it is of paramount interest to explore alternative technologies and architectures for the post-Moore's law era of computing. This project aims to build an integrated photonics computing system (from device to architectures) based on the residue number system (RNS) to achieve orders of magnitude improvements in computational speed per watt over the current state-of-the-art. The objectives of this research project are to:

• Offer potential transformative insights by exploring new materials for strong enhancements of light-matter-interactions.
• Explore attojoule per bit efficient and GHz-fast optical switching devices.
• Design and demonstrate compact 2×2 switches that are basic building blocks for optical residue arithmetic functions.
• Enable a novel approach to the design and evaluation of an entire class of optical compute engines based on residue arithmetic leading to multi-purpose computing.
• Explore co-design principles that relate device technology to the switch, the network architecture and the routing algorithm and methodology.
• Emulate and evaluate performance and accuracy using well-accepted community benchmarks.
• Pave the way for rapid and agile prototyping by enabling insights for advanced manufacturing on a silicon photonics platform.
• Enable the collective synergistic experience of the PI's, who are well established in their fields, to explore innovative nanophotonic computing paradigms.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	UU	37

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
GERNOT POMRENKE	426-8426

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

RANC: A Residue Arithmetic Nanophotonic Computer
Final Report (08/01/2019 – 12/31/2022)
AFOSR
FA9550-19-1-0277
PM: Dr. Gernot Pomrenke

Prof. Tarek El-Ghazawi (P.I.), tarek@gwu.edu, The George Washington University
Prof. Volker Sorger (co-P.I.), sorger@gwu.edu, The George Washington University

1. Accomplishments

1.1 Research Objectives

Due to the end of Moore's law and Dennard scaling, feature reduction and higher speed of clocking are seizing to be the source for higher computer performance. Therefore, it is of paramount interest to explore alternative technologies and architectures for the post-Moore's law era of computing. This project aims to build an integrated photonics computing system (from device to architectures) based on the residue number system (RNS) to achieve orders of magnitude improvements in computational speed per watt over the current state-of-the-art. The objectives of this research project are to:

- Offer potential transformative insights by exploring new materials for strong enhancements of light-matter-interactions.
- Explore attojoule per bit efficient and GHz-fast optical switching devices.
- Design and demonstrate compact 2×2 switches that are basic building blocks for optical residue arithmetic functions.
- Enable a novel approach to the design and evaluation of an entire class of optical compute engines based on residue arithmetic leading to multi-purpose computing.
- Explore co-design principles that relate device technology to the switch, the network architecture and the routing algorithm and methodology.
- Emulate and evaluate performance and accuracy using well-accepted community benchmarks.
- Pave the way for rapid and agile prototyping by enabling insights for advanced manufacturing on a silicon photonics platform.
- Enable the collective synergistic experience of the PI's, who are well established in their fields, to explore innovative nanophotonic computing paradigms.

1.2 Details of Accomplishments

During the first year of the project (08/01/2019 to 07/31/2020), our research is centered on the following three tasks:
• Design and simulation of efficient arithmetic RNS architectures, including RNS adders and multipliers, based on our previous introduced nanophotonic switch

• Design and simulation of the building blocks of an RNS convolutional neural network (CNN) accelerator architecture, such as the matrix-vector multiplication unit, activation unit, and pooling unit.
• Development of new programmable nanophotonic nonvolatile switches

We explored using hybrid plasmonic-photonic (HPP) switches to build different high-speed architectures of RNS computational units based on multistage interconnection networks. We studied the tradeoffs between the area and the control complexity of five different architectures for the residue adders and multipliers. In addition, we started to conceive the building blocks for the first photonic residue convolutional neural network accelerator, including the matrix-vector multiplication unit, activation unit, and pooling unit. Furthermore, we developed a new programmable nanophotonic nonvolatile switch based on phase change materials (PCM). Phase change materials have several unique properties, allowing them to become a very promising material in on-chip tunable optical devices.

During the second year of the project (08/01/2020 to 07/31/2021), our research focuses on the following four tasks:

- Design and simulation of efficient architectures of converters, including electronic binary to nanophotonic RNS convertor and nanophotonic RNS to electronic binary converter
- Design, simulation, and improvement of a system-level wavelength-division-multiplexing (WDM) enabled RNS convolutional neural network (CNN) accelerator architecture, deploying a pipeline in a layer-by-layer fashion with enhanced computational kernels
- Design of an RNS structured network architecture for the manycore system to integrate residue computing nodes, which addresses issues related to synchronization and collective operations.
- Design, fabrication, and measurement of a silicon-based photonic integrated circuit (PIC) of an RNS adder

We further explored the residue building blocks design. We developed an efficient architecture to convert electronic binary to nanophotonic RNS and convert nanophotonic RNS to electronic binary. We continued to work on the nanophotonic based residue CNN accelerator with wavelength-division-multiplexing (WDM) feature and optimized the design. We developed a space-efficient architecture to deploy the neural network pipeline in a layer-by-layer fashion and improved the sigmoid activation function units. The simulation results showed that the enhanced accelerator executed 9.8x faster and consumed energy 6.3x lower on average for the tested real machine learning benchmarks than the original design. Also, we designed an RNS structured network architecture for the manycore system, which integrated residue computing nodes while addressed synchronization and collective operations issues. Furthermore, we designed, fabricated, and measured a silicon-based photonic integrated circuit (PIC) of a passive modulo-4 adder. The PIC was demonstrated to work as expected in broadband (>30nm) and had a high signal-to-noise ratio (>5dB).

During the third year of the project (08/01/2021 to 12/31/2022), we needed to extend the project for five additional months due to COVID-19, as the closure of the facilities and subsequent delay of the fabrication company. Our research is centered on the following

three tasks:
• Design, simulation, and optimization of an RNS convolutional neural network (CNN) accelerator architecture with optimized residue computing blocks and evaluated it in a comprehensive comparison with other state-of-art accelerators, conducting a new space exploration study and application benchmarking.
• Design and simulation of an RNS structured network architectures in different control techniques for multicast operations in many-core system.
• Design, fabrication, and measurement of a wavelength-division-multiplexing (WDM) enabled silicon-based RNS CNN accelerator at the chip level.

The focus of the project of the third year was on optimizing the deep neural networks (DNN) accelerator with various enhancements, i.e., the residue matrix-vector multiplication unit, activation function unit, pipeline design, and comprehensive comparison with other accelerators and application benchmarking. The proposed design was evaluated using different CNN models. Given a similar power budget as the NVIDIA Tesla V100 and T4 GPUs, the proposed design operated on average more than 80x and 72x faster, respectively. We further analyzed the overhead of the conversions between binary and residue number system on application benchmarks. On average, the conversions took less than 2.34% of total execution time and 1.51% of total energy consumption. In addition, we designed and simulated two RNS structured network architectures for the manycore system. The optimal network had a bandwidth of 41 GB/s, which is 14.6x faster than the one of Blue Gene supercomputer. Moreover, we designed, fabricated and measured a WDM enabled silicon-based RNS matrix-vector multiplication (R-MVM) unit at the chip level with off-chip lasers and photo-detectors. The broadband PIC design demonstrated the functionality of a 4-bit R-MVM unit, which worked as expected in multi-wavelength with a high signal-to-noise ratio (>7dB).

1.3 Results Disseminated to Communities of Interest

The research team made significant efforts to disseminate the results of their project to the communities of interest. We presented their work at various international conferences, such as the IEEE Research and Applications of Photonics in Defense Conference, IEEE International Conference on Rebooting Computing (ICRC), International Conference on Parallel Processing (ICPP) and so on. We also published their findings in a high-impact peer-reviewed journal, ACM Journal on Emerging Technologies in Computing Systems. In addition, we submitted a patent application for the RNS adder. The dissemination efforts helped to increase the visibility of the project and its results, and facilitated collaborations with researchers and industry partners in the field.

The list of peer-reviewed Publications from this project is shown as follows:

Journal and Proceedings

1) Peng,J., Sun, S., Narayana,V.K., El-Ghazawi, T., & Sorger, V.J., "*Silicon Photonic Enabled Residue Number System Adder and Multiplier*" IEEE RAPID 2019
2) Peng, J., Alkabani, Y., Sun, S., Sorger, V. J., & El-Ghazawi, T. (2019, November).

Integrated Photonics Architectures for Residue Number System Computations. In *2019 IEEE International Conference on Rebooting Computing (ICRC)* (pp. 1-9). IEEE.

3) Peng, J., Alkabani, Y., Sun, S., Sorger, V. J., & El-Ghazawi, T. (2020, August). DNNARA: A Deep Neural Network Accelerator using Residue Arithmetic and Integrated Photonics. In *49th International Conference on Parallel Processing-ICPP* (pp. 1-11).
4) Peng, J., Sun, S., Narayana, V.K., El-Ghazawi, T. and Sorger, V.J., 2019, August. Silicon Photonic Enabled Residue Number System Adder and Multiplier. In *2019 IEEE Research and Applications of Photonics in Defense Conference (RAPID)* (pp. 1-2). IEEE.
5) Sun, S., Peng, J., El-Ghazawi, T. and Sorger, V.J., 2019, July. Nanophotonics Based Residue Number System. In Photonic Networks and Devices (pp. NeM3D-4). Optica Publishing Group.
6) Peng, J., Sun, S., Narayana, V., El-Ghazawi, T. and Sorger, V., 2019, July. Integrated Nanophotonics Enabled Residue Number System (RNS) Arithmetic. In 2019 IEEE Photonics Society Summer Topical Meeting Series (SUM) (pp. 1-2). IEEE.
7) Peng, J., Alkabani, Y., Puri, K., Ma, X., Sorger, V. and El-Ghazawi, T., 2022. A Deep Neural Network Accelerator using Residue Arithmetic in a Hybrid Optoelectronic System. ACM Journal on Emerging Technologies in Computing Systems (JETC), 18(4), pp.1-26.

- **Patent**

1) Jiaxin, P. E. N. G., Tarek El-Ghazawi, Volker J. Sorger, and Shuai Sun. "Residue arithmetic nanophotonic system." U.S. Patent Application 16/284,762. (Submitted)

2. Impacts

The project aimed to explore alternative technologies and architectures for post-Moore's law computing by building an integrated photonics computing system based on the residue number system (RNS). The research team focused on following tasks: designing integrated photonics residue arithmetic building blocks based on 2x2 optical switches, designing an RNS convolutional neural network (CNN) accelerator architecture, designing an RNS structured network architecture for the manycore system, and designing a WDM-enabled silicon-based RNS CNN accelerator at the chip level.

During the project, we explored the use of hybrid plasmonic-photonic (HPP) switches to build high-speed architectures of RNS computational units based on multistage interconnection networks. We further explored residue building blocks design and developed an efficient architecture to convert electronic binary to nanophotonic RNS and vice versa. A system-level CNN accelerator architecture was developed and demonstrated how to implement different CNN computational kernels using wavelength-division-multiplexing (WDM) enabled RNS-based integrated photonics. We also fabricated a WDM enabled silicon-based RNS matrix-vector multiplication (R-MVM) unit at the chip level with off-chip lasers and photo-detectors.

In addition, the team designed an RNS structured network architecture for the manycore system, which integrated residue computing nodes and addressed synchronization and collective operations issues. The design could lead to multi-purpose computing and rapid and agile prototyping with enabling insights for advanced manufacturing on a silicon photonics platform.

Moreover, one of the major contributions of our project was the development of a space and energy-efficient CNN accelerator. This accelerator has the potential to be applied in edge computing, where space and energy are limited, and on the Internet of Things (IoT), where devices need to perform complex tasks with minimal power consumption. Furthermore, our project extended the existing techniques and methodologies for CNN inference. These techniques have the potential to advance state-of-the-art in deep learning and computer vision, which are rapidly growing fields with applications in a wide range of industries.

Overall, the project's contributions to the field of post-Moore's law computing include innovative nanophotonic computing paradigms, efficient residue building blocks design, and advanced CNN accelerator and network architectures. These contributions could lead to transformative insights and significant improvements in computational speed per watt, paving the way for the post-Moore's law era of computing.

3. Changes

As a result of the COVID-19 pandemic, we found it necessary to extend the project by a period of five months in the third year, due to the closure of facilities and the subsequent delay of the fabrication company.

In our project, we worked on the design of Wavelength Division Multiplexing (WDM)-enabled photonic residue computing blocks, which utilize the properties of light for high-speed computation. As we developed this technology, we discovered that the high-parallelism design of our blocks could be particularly suitable for neural networks. We subsequently conducted further research and proposed a chip-level design that included a matrix-vector multiplication unit, activation function unit, and conversion units between the residue number system and binary system to make it an end-to-end system. One of the key advantages of our proposed design was its speed. We tested our design under similar power budgets as the NVIDIA Tesla V100 and T4 GPUs, and found that it operated on average more than 80x and 72x faster, respectively. This demonstrated the potential for photonic computing to significantly outperform traditional computing methods in certain applications.

4. Technical Updates

Here we discuss in further detail the research carried out and results obtained yearly.

4.1 Year 1

4.1.1 Efficient RNS arithmetic Architecture

4.1.1.1 RNS Adders

We developed three novel architectures for RNS computational units, including shifting RNS adder, Benes RNS adder, and Arbitrary Size Benes (AS-Benes) adder. They were built based on our prior introduced indium tin oxide ITO 2x2 switch [1]. We compared the newly proposed architectures to the ones previously proposed, including Tai RNS adder [2] and all-to-all sparse directional (ASD) adder [3]. We showed the tradeoff between the area savings and the complexity of the control unit and published the results in [4].

The fundamental component for the integrated photonics RNS arithmetic device is the 2x2 switch with two states that are controlled by an electrical signal, shown in Fig. 1 The two states, bar state and *cross* state, could be switched by the control signal. In the *bar* state, the input light source will propagate straight forward through the switch (Fig. 1(a)), whereas in the cross state, the light source will be routed to the opposite output (Fig. 1(b)). A voltage control signal will be applied to the switch to set its state. We chose the hybrid plasmonic-photonic (HPP) ITO switch [1] to implement our RNS adder due to its compact size, high speed, and low power consumption.

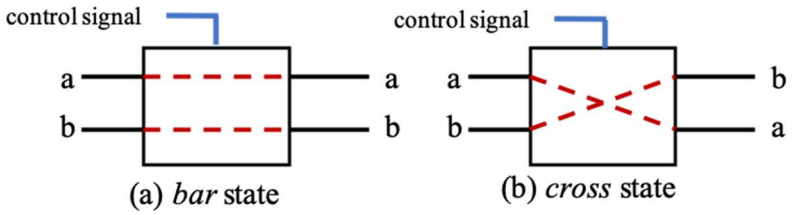


Fig. 1. Conceptual Schematic of Two States of a 2x2 Switch. (a): *bar* state. (b): *cross* state.

The adder is one of the most critical computational units for an RNS arithmetic system. The RNS adder based on the photonic 2x2 switch is considered to be the basic computation block since other computational units could be implemented using the adders, including subtraction, multiplication, and evaluation of polynomials. A residue adder can be implemented based on the shifting property of a residue additive operation or the adder's routing capability when considering it as a network.

We discussed five architectures of RNS adders are shown as Fig. 2. Two of those architectures (Tai and ASD) have been previously proposed to be used as residue adders. The residue additive operation acts as a simple shift right operation when one-hot-encoding is deployed. Tai et al., proposed an RNS adder with an electro-optical component in work [2]. We devised the shifting RNS adder, which is an improved architecture based on the Tai RNS adder. Both two architectures are implemented using the shifting properties of RNS addition. While these architectures' control logic is simple, they require many optical switches, scaling to $O(N^2)$ for a modulo-N system. The RNS adder could be considered as a routing network as well. The all-to-all sparse directional (ASD) RNS is proposed as a

non-blocking all-to-all communication network [3, 5]. To reduce the overhead of optical components, we propose a routing network based on Clos and Benes networks [6] or Arbitrary Size Benes (AS-Benes) networks [7]. When the number of inputs of AS-Benes equals 2^n (n is an integer), it will be the same as the Clos and Benes network. Thus, we combine these two architectures in the same figure.

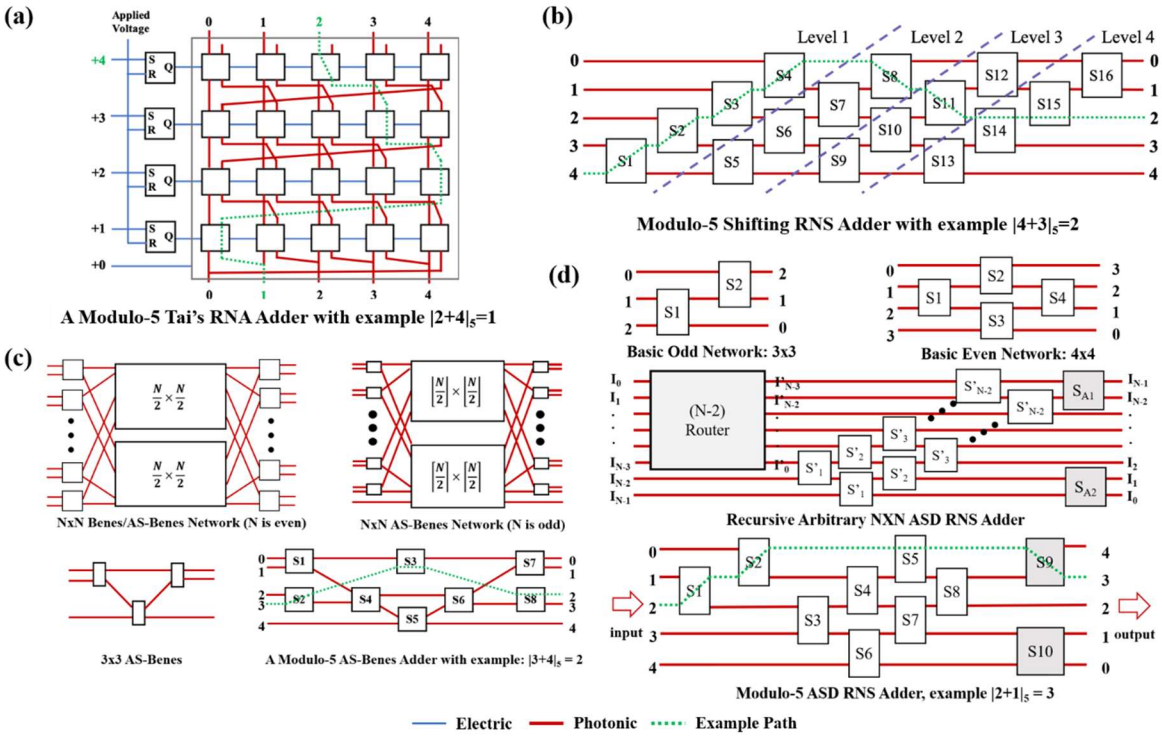


Fig. 2. Architectures of an RNS Adders. (a) A Tai's RNS Adder. (b) A Shifting RNS Adder. (c) A Benes/AS-Benes RNS adder. (d). An ASD RNS Adder.

Table I shows the tradeoff of a Modulo-N RNS adder with different architectures. The Tai and shifting RNS adders are designed by the shifting property for the residue additive operation and controlled by a logic circuit. The designs have high scalability; however, the requirements for the optical components are large with complexity $O(N^2)$. The Tai model utilizes the greatest number of switches; however, the control logic is easy to implement, and the light propagation delay is less than the shifting adder.

Alternatively, the routing-based RNS adders, including ASD, Benes, and AS-Benes, require fewer optical components. Nevertheless, the tradeoff is the complexity of the control unit implementation. Since the control bits are stored in a LUT, the area of the LUT and the time of reading the LUT are additional overheads in this case.

We compared the newly proposed architectures to the ones previously proposed and showed the tradeoff between the area savings and the complexity of the control unit. Moreover, we considered practical implementations of 32, 64, and 128-bit number systems using those architectures and showed that most of the prior architectures could not be used

to implement such systems except for our newly proposed computational unit based on arbitrary size Benes networks. Evaluation of the area and delay of all networks showed that the AS-Benes-based computational unit saves up to 90% of the area and is up to 16 times faster than the other architectures when considering the photon time-of-flight.

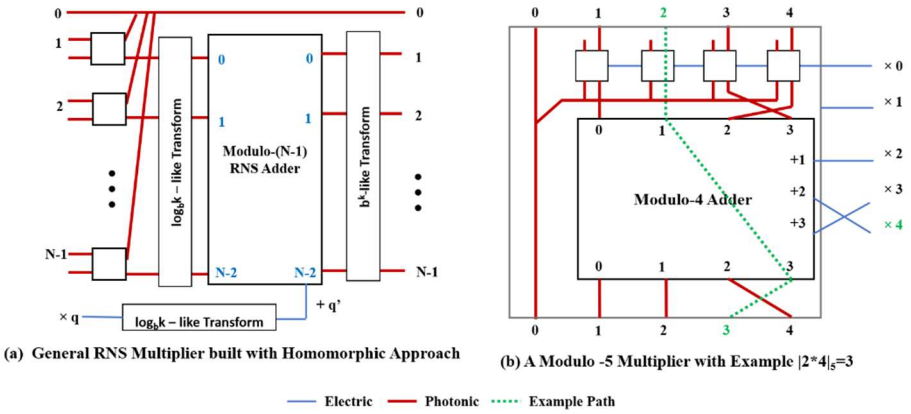
Table I. Comparison of a Modulo-N RNS adder built on different architectures, including Tai, shifting, all-to-all sparse directional (ASD), Benes, and Arbitrary Size Benes (AS-Benes).

Parameters	Tai	Shifting	ASD	Benes	AS-Benes
# of switches S(N)	N(N-1)	$(N-1)^2$	Odd: $(N-1)^2/2+2$ Even: $N(N-2)/2+2$ (N>2)	$N \log_2 N - N/2$	$S(N) = 2\left\lfloor\frac{N}{2}\right\rfloor + S\left(\left\lfloor\frac{N}{2}\right\rfloor\right) + S(\left\lfloor\frac{N}{2}\right\rfloor)$ S(1) = 0; S(2) = 1
Control Unit Implementation	Logic Circuit		LUT		
# of Ctrl. Bits	N-1		S(N)		
Size of LUT (bit)	-		N*S(N)		
Max # of Stages	N-1	3N – 6 (N>=4))	ST(N) = ST'(N-2) + 4 ST'(3) = 2, ST'(4) = 3 ST(2) = 1, ST(3) = 3, ST(4) =4	$2\lceil log_2 N \rceil - 1$	

4.1.1.2. RNS Multiplier

Residue multipliers' design is different from the residue adder design due to the zero-factor, and thus it could be decomposed into two parts. The first one contains the situation that there is at least one zero-factor in the operation. All input light signals should be routed to the position 0. Hence, this path should be built from not only the input position 0 but also any other input positions. The 2x2 switch helps make the selection.

The other part is an (N-1)-to-(N-1) RNS with position range [1, N-1] (Fig. 3 (a)). This part needs to be designed to route the input light according to the results of the multiplication operation. It is straightforward to implement such a schematic with a routing-based modulo-(N-1) adder. All the considered routing based RNS adders have been proven to work well as a rearrangeable non-blocking network. This means that a network can realize



(a) General RNS Multiplier built with Homomorphic Approach

(b) A Modulo -5 Multiplier with Example $|2*4|_5=3$

Fig. 3. Architectures of an RNS Multiplier. (a): A Modulo-(N-1) RNS multiplier implementation using homomorphic approach. (b): A modulo-5 RNS multiplier.

all possible permutations between inputs and outputs, requiring rearranging the existing connections. Hence, it is sufficient to build a routing based RNS multiplier by generating another LUT storing the switch states required for multiplication results.

Alternatively, the homomorphic approach could be applied to the RNS multiplier. A modulo-N multiplier could be converted to a $log_b\, k$-like transform, a modulo-(N-1) additive operation, and a b^k-like transform. If either multiplicand for the multiplication operation is 0, the product will be 0. Fig. 3 (b) shows the transform for a modulo-5 multiplier.

By transforming both X and Y in operation $X*Y$, the modulo-N multiplication could be converted to the modulo-(N-1) addition. This homomorphic approach could be applied to any RNS adder architecture. Comparing to the conventional multiplication in binary systems, the other advantage for RNS multiplication is that it could reach the same latency as RNS addition.

4.1.2 An RNS convolutional neural network accelerator architecture

We further introduced a novel DNN accelerator using residue arithmetic (DNNARA) based on integrated photonics and published our work in International Conference on Parallel Processing (ICPP) 2020 [8]. DNNARA attempts to provide a practical accelerator from both the area and the power. The use of RNS enables us to reduce the optical critical path to maintain low laser power. Moreover, the one-hot-encoding used to encode the numbers enhances the speed of switching between the optical and electronic domains, enabling our system to have high throughput. We took advantage of WDM to achieve a high level of parallelism while maintaining a reasonable area. Our residue Matrix-Vector Multiplication module can achieve two orders of magnitude higher throughput when compared to the memristor crossbar. When using the same power budget as the NVIDIA Tesla V100 GPU, our chips can make inference more than 19.3x faster.

4.1.2.1. Overview

We attempted to reduce our chip area while maintaining a high level of parallelism using WDM-enabled MACs combined with filters, as illustrated in block 3 in Fig. 4. In this example, we executed the three independent MACs using one optical MAC, where each row is input to the same MAC using different frequencies. At the output, different results are extracted using filters. In order to have a power-efficient chip while maintaining scalability, we aimed to reduce the optical critical path by using an RNS-based implementation of our system. In RNS, a number is represented as a set of smaller numbers (residue set) for pairwise co-prime moduli. It exposes the high potential for parallelism because residue arithmetic is digit-irrelevant for both addition and multiplication. For each residue, all operations could be executed independently of the other residues, and the results only need to be combined at the end.

The use of RNS computations has two main advantages: (1) using smaller optical MACs with shorter critical paths that are power efficient as the optical losses are minimized through the circuit, and (2) numbers in our system are represented using one-hot encoding, and this relieves the need of using DACs/ADCs.

The proposed accelerator, DNNARA, is built based on residue arithmetic and integrated photonics, as shown in Fig. 5. The proposed DNNARA chip is built with several tiles, and each tile consists of several residue MVM (R-MVM) units. Furthermore, it comes with additional hardware support, including binary-to-residue (bin2RNS) converters, residue-to-binary (RNS2bin) converters, activation function units, max pooling units, LUTs that store the switch states and input/output information, eDRAMs that store the data information, one additional R-Accumulator unit that could accumulate the results from different tiles or R-MVM units, as well as a bus that connects all the units.

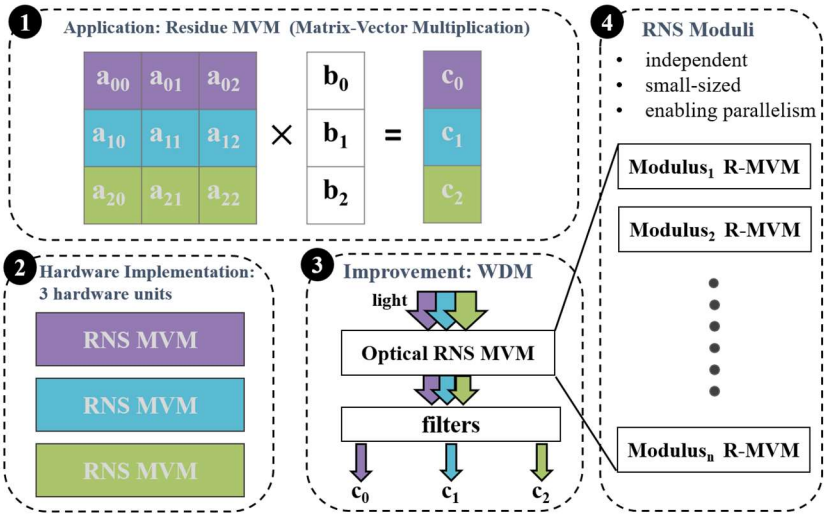


Fig. 4. Matrix-Vector Multiplication (MVM) using WDM-enabled Optical RNS MAC.

4.1.2.2. Integrated Photonic Residue Arithmetic Computing Engine for Neural Network

The proposed R-MVM, max-pooling unit, and activation function unit are illustrated in Fig. 6. A residue MVM (R-MVM) unit executes the MAC operations, which could be implemented with R-Adders and R-Multipliers. Besides, the wavelength division multiplexing (WDM) capable of selected photonic devices allows several operations executing simultaneously. The following example shows how an R-MVM unit performs

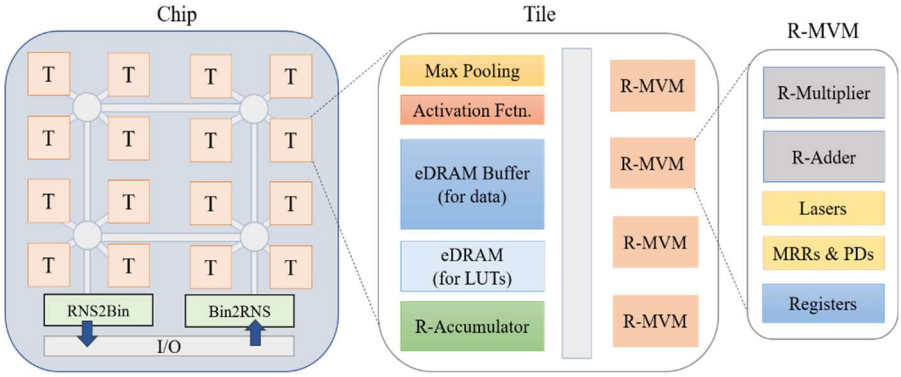


Fig. 5. Overview of a DNNARA chip.

several MAC operations in a neural network using WDM.

Assuming there are a 5x5x1 input feature and a 2x2 kernel, with stride size one and zero padding (Fig. 6(a)). The convolution of the first two columns of the input feature with the kernel will be mapped to a matrix-vector multiplication and addition (MVMA), $Y = X * W + B$. The matrix $\boldsymbol{X}$ is not the same as the input feature. It is rearranged according to the kernel size. Vector $\boldsymbol{B}$ represents the bias value. Vector $\boldsymbol{W}$ is set as the electrical control signal for the multipliers, and matrix $\boldsymbol{X}$ will be encoded as an optical signal. For instance, w_0 will be treated as the select signal to load the corresponding switch states from the LUT. The R-Multiplier will be set up accordingly.

Then the first row of matrix X will be encoded as a light signal for four R-Multipliers. By utilizing the WDM feature, the first column of X (x_0, x_1, x_2, x_3) will be represented by different wavelengths, $\lambda_0, \lambda_1, \lambda_2, \lambda_3$, respectively. They will be injected into corresponding input ports. Several sets of micro-ring resonators (MRRs) filter out the light with designated wavelengths.

The photo-detector (PD) set after each MRR will detect the light. If the light passes through the same waveguide with other wavelengths, it will propagate through this MRR with low loss until it reaches the MRR with designated resonance frequency. For instance, the first

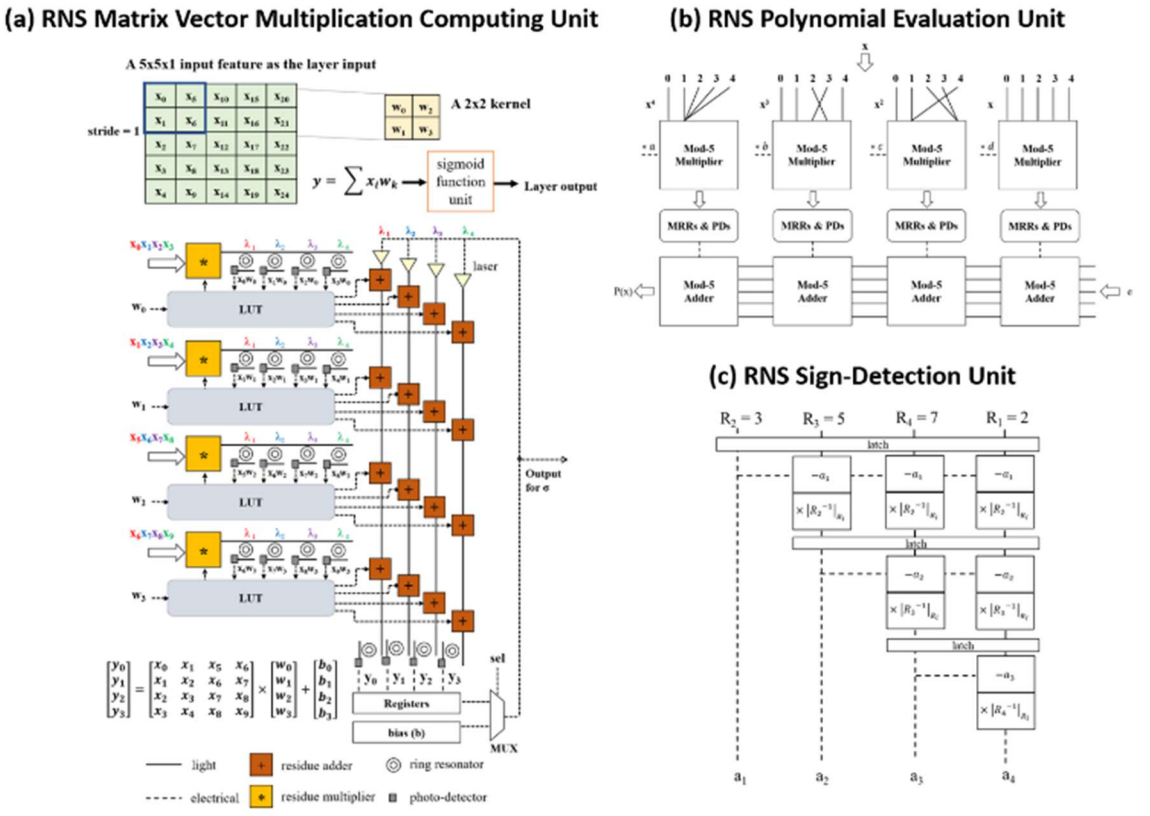


Fig. 6. RNS Computing Blocks of Proposed CNN Accelerator. (a). An RNS Matrix-Vector Multiplication Unit. (b). An RNS Polynomial Evaluation Unit. (c). An RNS Sign-Detection Unit.

MRR is designed for λ_0, and the PD will detect only the light at the same wavelength. Therefore, based on the signal of the PD, the result of the operation would be derived. Note that each row will be encoded as the same wavelength for all the multipliers, decreasing the hardware's complexity. After detecting the multiplication results, optical signals would be converted to electrical signals and set the states for the switches inside a group of residue adders. These sets of adders will sum the multiplication results. Light will be injected again based on the bias in this case. Then the result, y_i, would be stored in registers.

The size of matrix X should be more extensive in real applications. Due to the area/power constraints, there will not be unlimited resources in an R-MVM. To further explore the opportunity of a residue MAC operation for large-sized vector, matrix-vector multiplication is split into smaller sizes, which could fit in a single R-MVM unit. Then, sub-operations could be executed in other R-MVM units before accumulating all multiplied results. Registers would be required to store the partial results. The R-Accumulator in the tile will accumulate the partial sum that is distributed into different R-MVMs.

Moreover, it is hard to support too many wavelengths in such a system. We limited our design to a small number of frequencies (WDM size), 3~5, resulting in the necessity to decompose of the X. Hence, the register at the end of each adder does not always store the result of such operations. It may store the partial sum instead. A multiplexer (MUX) would select either the partial results or the bias. Lasers inject light based on the selection.

The sigmoid function is popular as activation functions in deep neural networks and is designed in the proposed work. However, floating-point calculations exist in such function. Hence, the Taylor series, expressed as a polynomial, is selected to express the sigmoid function, which contains residue adders and multipliers only, as illustrated in Fig. 6 (b). The hardware requires more attention in design time since each term for the power of x should be processed in advance, and specific hardware would be fabricated accordingly. In addition, if the activation function is always the same type, then the switch requires one-time initialization only because the coefficient in the Taylor series for one specific function is always the same.

A residue max-pooling unit is designed to reduce the conversions between the binary and the residue domain. In order to find the maximum value, first, a subtractor is needed, which could be easily developed by the residue adder since negative notation is available in residue arithmetic. A sign detector would be useful to identify the larger number according to the result of the subtraction. The sign in the RNS is implicit, which means the sign is part of the number representation itself. Thus, it is not straightforward to determine if a residue representation is positive or negative. In conventional number systems, the sign representation is explicit. The two's complement representation of the binary system, for example, utilizes the most significant bit to represent the sign itself. We adopted a mixed-radix conversion process for residue sign detection, which involves subtraction and multiplication only. The architecture is shown in Fig. 6 (c).

4.1.2.3 Evaluation

We performed design space exploration to select efficient architecture parameters. We

chose the metric of the *computational capability*, which represents the number of 16-bit operations per second per mm^2 per Watt ($GOPs/s \cdot mm^2 \cdot W$). This metric helps us to evaluate the whole system in a more general way, considering the computing throughput, area, and energy at the same time. Our design reaches 12.6 ($GOPs/s \cdot mm^2 \cdot W$), with *5 WDM, 32 R-MVMs in a tile, and 8 tiles in a chip* as the configuration. Compared to memristor crossbars, our residue matrix-vector multiplication unit has two orders of magnitude higher peak performance.

In addition, we devised a system-level simulator with the optimal configuration. We apply our architecture on modern neural network architectures for further performance evaluation. We adopted the modified VGGs, DeepFace, ResNets, as well as LeNet-5. NVIDIA reports that one of the most recent GPUs, Tesla V100, could perform one image in 1.2 ms when running VGG-4 while the power consumption is 250 Watt. The batch size equals one for inference in this scenario. With a similar power budget, 12 DNNARA chips could be used, and one image could be processed in 0.062 ms in such a system, achieving a speedup of 19.3x.

4.1.3 Nonvolatile Phase-Change Directional Coupler Switch

Phase change materials (PCM) have several unique properties, allowing it become a very promising material in on-chip tunable optical devices: the phase transition between amorphous and crystalline states leading to the significant change of optical constant, state retention without extra power, fast and reversible switching of the state of microsecond light or electric pulse, and excellent scalability. The material Germanium-antimony-selenium (GSSe) has a very low loss in the amorphous state and considerable refractive index change, allowing us to achieve a low-loss optical switch. It makes GSSe suitable for photonic integrated circuits (PICs), enabling a large (NxN) PIC switching fabric, for example, for RANC.

Here, we demonstrated a compact (~40 μm), low-loss (~1 dB), and 2×2 switches using the PCM, Germanium-antimony-selenium (GSSe), based on the nonvolatile programmable GSSe-on silicon structure, the asymmetric directional coupler (DC) switch, and a bypass heater design.

Microscope picture shows the structure of the 2 × 2 DC switch. The asymmetric coupling region consists of a regular silicon strip waveguide, and a GSSe covered silicon hybrid waveguide, where a thin layer of GSSe is placed on silicon. The heater was designed to be in close proximity to the hybrid waveguide and is used to control the state of GSSe via thermo-electric heating.

To determine the best design parameters, we used the finite-element method (COMSOL Multiphysics) to analyze the effective index of the eigenmodes supported in the normal silicon waveguide and the GSSe covered silicon hybrid waveguide. The width of the silicon strip waveguide is chosen as 480 nm to ensure single-mode operation. To match the phase, the width of the hybrid waveguide is chosen as 420 nm. The gap between the two waveguides is chosen as 200 nm, which is considered the tradeoff between the insertion loss and the coupling length. The normalized optical field intensity distribution of the

switch and the transmission are studied by 3D finite-difference time-domain method (Lumerical). The difference of transmission in bar state and cross state both are above 6 dB.

The device is based on an SOI platform, which is taped out from the Applied Nanotools Inc. After one EBL step, 20 nm GSSe and 10 nm silicon dioxide were deposited on the top of the silicon waveguide. The tungsten heater was completed using the second EBL step followed the sputtering and lift-off process. The microscope picture shows the fabricated 2 × 2 DC switch.

The spectral response of the device was measured by the off-chip fiber system. For each device, the initial state of PCM is at the amorphous state because of thermal deposition. We measured the transmission of the output port. After that, a series pulse with 10 V and 1 us width were sent to the heater, which allows the heater to reach the temperature of the phase transition from aGSSe to cGSSe (600 K). The measurement results of DC switch show that for the wavelength range 1540 to 1560 nm, the extinction ratio is about 12-13 dB, and the insertion loss is relatively low (~1 dB).

4.2 Year 2

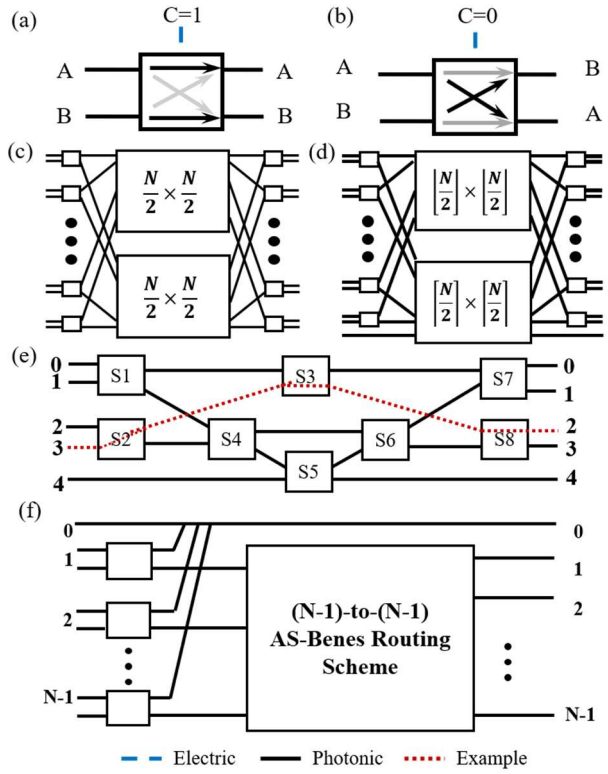


Fig. 7. Schematic of Residue Adders and Multipliers. An electro-optical 2x2 switch in (a) bar state and (b) cross state. Schematic of an AS-Benes residue adder if the input is (c) even or (d) odd. (e) An AS-Benes modulo-5 adder. (f) A Modulo-N Residue Multiplier Implementation.

4.2.1 Efficient RNS arithmetic Architecture

4.2.1.1 Residue Adders and Multipliers

Our previous research studied five different residue adder architectures, including Tai RNS adder [2], all-to-all sparse directional (ASD) RNS adder [3], shifting RNS adder, Benes RNS adder, and Arbitrary Size Benes (AS-Benes) adder [4]. All adders were built based on our prior introduced indium tin oxide ITO 2x2 switch [1]. Our study showed that AS-Benes residue adder/multiplier gained the highest performance in terms of area and power. Hence, the AS-Benes network residue adder/multiplier was selected in our work. The designs are shown in Fig. 7

4.2.1.2. Binary to RNS Converter

Efficient conversion between binary and one-hot RNS representation is essential if we integrate this unit as part of a digital computer. Here, we propose the architecture of the electronic binary to nanophotonic RNS converter as illustrated in Fig. 8, as inspired by [1]. A binary number X could be converted to a decimal number by summing up all its digits times the corresponding position weight, e.g., the power of two. It could be expressed as $X = \sum b_i * 2^{i-1}$, where b_i represents the number of *i-th* digit in the binary representation. Congruences with respect to the modulo-m are satisfied in residue addition, and thus, the residue representation of X is $X = |\sum b_i \times |2^{i-1}|_m|_m$. The representation of each weight in the binary system with respect to the modulo-m system needs to be pre-calculated at the design time. Several electronic multiplexers are applied to each nanophotonic residue adder, choosing the summand according to the binary representation. A laser source is placed at the input port zero for the accumulation. If b_i is zero, the summand will be zero; otherwise, it will be $|2^{i-1}|_m$, which is pre-calculated at the design time. Finally, the photodetectors at the output ports will detect lights, and the residue representation will be derived based on its position. The simulation result shows that one binary-to-RNS converter is 0.06 mm^2 and consumes 35 mW, while the electronic parts use 32nm technology.

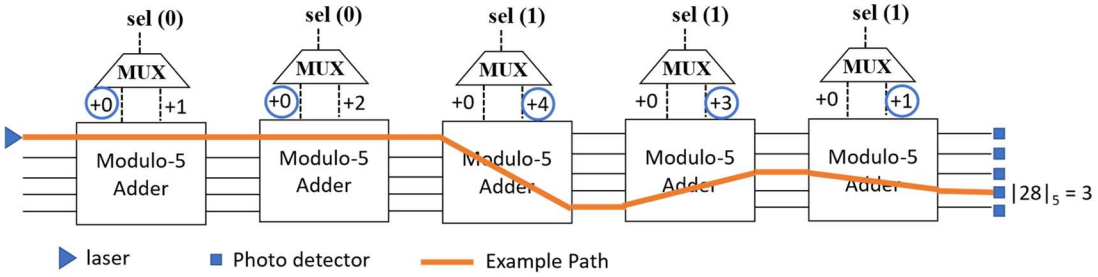


Fig. 8. Binary to RNS Converter. The example shows that number 28 in electronic binary, $(11100)_2$, will be converted to nanophotonic modulo-5 RNS ($|28|_5=3$).

4.2.1.3. RNS to Binary Converter

RNS to binary conversion is more complicated than the forward conversion (from binary to RNS). There are two basic classic methods proposed for the conversion from RNS representation to binary notation, including the Chinese Remainder Theorem (CRT) and the Mixed-Radix Conversion (MRC) based technique [9]. Other methods vary from these two, with some improvements according to the selected moduli set or the easy adaption for the chosen approach. The CRT-based conversion is straightforward, which would convert the number from RNS to binary directly. The MRC-based conversion needs twice conversions. First, the number will be converted to MRC, and then it will be further converted to a conventional notation. Both the CRT- and the MRC- based conversions contain addition, multiplication, and multiplicative inverse only, which could be implemented by previously introduced optical residue adders and multipliers. However, the accumulation process in the CRT-based conversion may require a larger number or modulus to derive the result, which is not suitable for the proposed nanophotonic devices. Thus, we choose the MRC-based conversion here.

Fig.9 shows the architecture of the RNS to MRC by utilizing integrated photonic residue adders and multipliers. Assume that the chosen moduli set is {3, 5, 7}, first the RNS representation will be converted to MRC, with coefficients {b_1, b_2, b_3}. Several general multipliers and adders will be applied after, performing $X=b1\times(R2\times R3) + b2\times (R3) + b3$ for the final result. As the number of modulus selected increases, the calculation time will increase due to the dependency between digits calculations. Fortunately, the proposed architecture could be pipelined [2]. Additional hardware support (i.e., the latches, more lasers, and photodetectors) are required. High throughput and conversion speed will be benefit from the pipeline design. It needs ten cycles to perform the first conversion for 16-bit data width with all prime moduli set. After that, each cycle would generate one more result. According to our simulation results, one RNS-to-MRC converter is 0.025 mm^2 and 2.4 mW, based on 32 nm technology for electronic components.

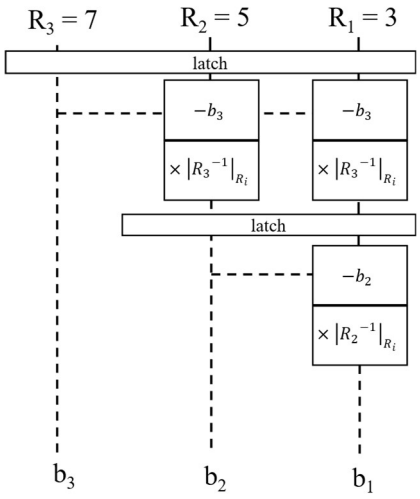


Fig. 9. Example Schematic of an RNS to MRC Converter.

4.2.2. Enhanced RNS CNN accelerator architecture

We introduced a novel CNN accelerator using residue arithmetic (DNNARA) based on integrated photonics and published our work at International Conference on Parallel Processing (ICPP) 2020 [6]. DNNARA attempted to provide a practical accelerator from both the area and the power. The use of RNS enables us to reduce the optical critical path to maintain low laser power. Moreover, the one-hot-encoding used to encode the numbers enhances the speed of switching between the optical and electronic domains, enabling our system to have high throughput. We took advantage of WDM to achieve a high level of parallelism while maintaining a reasonable area. To get better performance, we improved the design of the sigmoid activation function unit. In addition, we continued to work on the residue matrix-vector multiplication (R-MVM) unit. We developed a space-efficient architecture to deploy the neural network pipeline in a layer-by-layer fashion to perform the massive amount of multiply-accumulation (MAC) operations in a CNN, allowing higher throughput for the system. The simulation results indicated that the enhanced accelerator executed 9.8x faster and consumed energy 6.3x less on average for tested benchmarks than the original design. Furthermore, we run the pre-trained machine learning models on NVIDIA Tesla V100, one of the most recent GPUs. The power consumption of such a GPU was 250 Watts. The batch size equaled one for inference in this scenario. With a similar power budget, 12 DNNARA chips could be utilized, and one image could be processed 50.2x faster on average.

The sigmoid function is popular as an activation function in deep neural networks and is designed in the proposed work. However, floating-point calculations exist in such a function. Hence, the Taylor series, expressed as a polynomial, is selected to represent the sigmoid function, which contains residue adders and multipliers only (Fig. 6(b)).

Furthermore, we noticed that the Taylor series coefficients for the sigmoid function are always the same, resulting in the same accumulation results. Thus, we further improved the sigmoid function unit. Instead of using the several residue multipliers and adders, a mapping design that pre-calculates all possible results are proposed. For instance, a polynomial, $P(x) = ax^4 + bx^3 + cx^2 + dx + e$, where a=b=c=d=1, and e=0, then in a modulo-5 system, all five possible inputs have the corresponding results as follow: P(0) = 0; P(1) = 4; P(2) = P(3) = P(4) = 0, illustrated in Fig.10 (a). Moreover, this design benefits from the WDM feature, which allows multiple numbers to add the non-linearity simultaneously

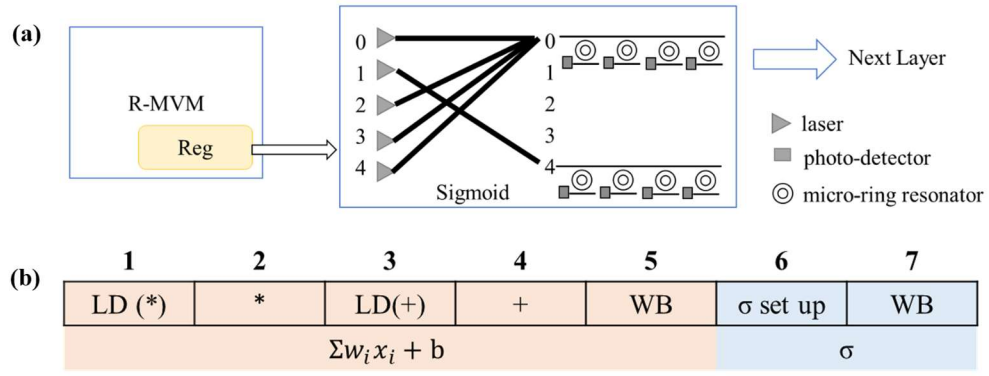


Fig. 10. (a) Schematic of Enhanced Sigmoid Activation Function Unit Design. (b)Proposed Pipeline Design of MAC Operation and Sigmoid Activation Function.

by using one hardware set. However, the hardware requires more attention in design time since the mapping should be processed in advance, and specific hardware would be fabricated accordingly.

To make our photonic system compatible with the electronic component and the conservative design for current technology, the frequency for the proposed system is 1.2GHz. At cycle 1, all the switches inside the four multipliers load the corresponding states from the LUT depend on the multiplicands. The reading time from the LUT needs around 0.5ns (simulation result from CACTI 7.0 [10]), which fits designated one cycle. The multiplicative operations are executed in optical residue computing engines at cycle 2. First, all the switches are set to desired states concurrently. The response time of the 2x2 switch is as low as 5.1ps. Then lasers inject lights to the four multipliers simultaneously. Light traverses at extremely high speed, and it costs less than 1ps for the residue computing engine in the worst case. MRR resonates light based on different wavelengths at around 130GHz, and finally, the PD detects light at 35GHz. Thus, from injection to detection, the overall light traveling time would be around 60ps, which is much shorter than the designed clock cycle. To guarantee that the optical switches are being set before the light is injected, a single cycle could be separated as two sub-cycles. The switch is set up in the first half cycle, and the light will be injected and detected in the second half cycle. Instead of executing the multiplications, cycle 3 and cycle 4 are designed for the additive operation, performing similarly to cycle 1 and cycle 2. In the end, results should be stored into the registers, consuming one more cycle (cycle 5). Therefore, 32 operations could be executed in 5 cycles in this example. It is worth noting that addition and multiplication are considered as two separate operations. By taking advantage of the pipeline design, 32 more operations will be fully executed each cycle after the first five cycles. However, it needs additional hardware support, including latches between stages, more lasers, and photodetectors.

The activation function is pipelined to gain more performance (Fig. 10 (b)). To perform a sigmoid function, we need to initialize the optical components, inject light, detect light, and finally store the result in the memory. Here, we always utilize the same number of wavelengths that an R-MVM can keep the compatibility. The first cycle of the activation function (cycle 6) is similar to cycle 2, which aims to initialize the optical activation blocks, inject and detect light. Light injected into the activation blocks will be routed to the expected location, decoded as the results. Finally, the results are written back to registers in cycle 7. Compared to the original design, the simulation result shows that the proposed sigmoid function architecture saves around 51.9% area and 25% power.

Evaluation. We performed design space exploration to select efficient architecture parameters. We chose the metric of the *computational capability*, which represents the number of 16-bit operations per second per mm^2 per Watt ($GOPs/s \cdot mm^2 \cdot W$). This metric helps us to evaluate the whole system in a more general way, considering the computing throughput, area, and energy at the same time. Our design reaches 12.6 ($GOPs/s \cdot mm^2 \cdot W$), with *5 WDM, 32 R-MVMs in a tile, and 8 tiles in a chip* as the configuration. Compared to memristor crossbars, our residue matrix-vector multiplication unit has two orders of magnitude higher peak performance.

In addition, we devised a system-level simulator with the optimal configuration. We apply

our architecture to modern neural network architectures for further performance evaluation. We adopted the modified VGGs, DeepFace, ResNets, as well as LeNet-5. We wrote a simulator for the proposed chip and ran all the benchmarks on it. Our simulation results indicate that the intra-layer pipeline accelerator gained an average of 9.8x faster and consumed energy an average of 6.3x less on tested benchmarks than the original design. In addition, we run the pre-trained models on NVIDIA Tesla V100, one of the most recent GPUs. The power consumption of such a GPU is 250 Watt. The batch size equals one for inference in this scenario. With a similar power budget, 12 DNNARA chips could be used, and one image could be processed 50.2x faster on average.

4.2.3 An RNS structured network architecture for the manycore system

We further propose an architecture of residue computing nodes in the manycore system, addressing collective operations and synchronizations issues. Collective operations are an integral part of parallel computing paradigms and involve all the nodes in the parallel program. Due to the participation of all nodes, reduction operations (such as addition, multiplication, min, and max operation) are expensive. Also, synchronization operations in large-scale systems can consume much power and incur performance penalties due to the need for all cores to communicate with each other [11-12]. One of the common synchronization operations is the *barrier*, which requires all the participating cores to stop execution and wait until all cores have arrived at the barrier before further executing the rest of the program.

Integrated photonic provides a viable means for integrating barriers within the communication network at very high performance. Our proposed 2×2 switches are particularly useful for barrier implementation. Fig.11 shows an example with a synchronization barrier and collective sum operation that accumulates all partial results from different cores. Each processor (P0-P3) controls one 2x2 switch, which sets the switch in either bar or cross state. Light cannot propagate through the switches unless all of them are set as *bar* states, which are available once the computation results are ready. Then lights are able to be routed to a set of residue adders, accumulating all results at high speed. For a 16-bit system, the area and power of the collective unit will be around $0.004*p$ um^2 and $0.8*p$ Watt, respectively, where p represents the number of processors in the

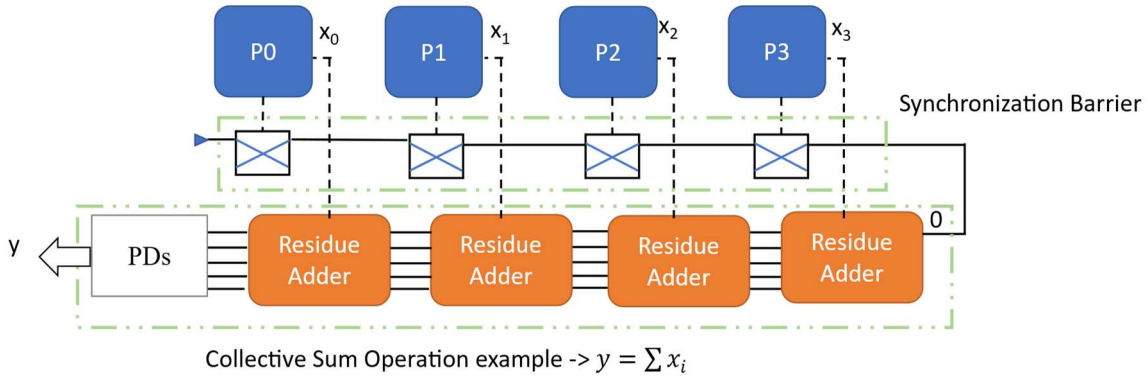


Fig. 11. Schematic of RNS Cores for Synchronization Barrier and Collective Sum Operation.

system. The addition operation is performed entirely in photonics, and the total sum appears in the output of the last core in the RNS format, in the optical domain. There are no intermediate electronic-optic-electronic conversions required before/after the addition operation at each core. Each addition is carried out on the fly along with the data routing. Once the inputs are set up, the time to completion depends entirely on the speed of light alone.

4.2.4 Nonvolatile Phase-Change Directional Coupler Switch

Phase change materials (PCM) have several unique properties, allowing it to become a very promising material in on-chip tunable optical devices: the phase transition between amorphous and crystalline states leading to the significant change of optical constant, state retention without extra power, fast and reversible switching of the state of microsecond light or electric pulse, and excellent scalability. The material Germanium-antimony-selenium (GSSe) has a very low loss in the amorphous state and considerable refractive index change, allowing us to achieve a low-loss optical switch. It makes GSSe suitable for photonic integrated circuits (PICs), enabling a large (NxN) PIC switching fabric, for example, for RANC.

Here, we demonstrated a compact (~40 μm), low-loss (~1 dB), and 2×2 switches using the PCM, Germanium-antimony-selenium (GSSe), based on the nonvolatile programmable GSSe-on silicon structure, the asymmetric directional coupler (DC) switch, and a bypass heater design.

The microscope picture shows the structure of the 2 × 2 DC switch. The asymmetric coupling region consists of a regular silicon strip waveguide and a GSSe covered silicon hybrid waveguide, where a thin layer of GSSe is placed on silicon. The heater was designed to be in close proximity to the hybrid waveguide and is used to control the state of GSSe via thermo-electric heating.

To determine the best design parameters, we used the finite-element method (COMSOL Multiphysics) to analyze the effective index of the eigenmodes supported in the normal silicon waveguide, and the GSSe covered silicon hybrid waveguide. The width of the silicon strip waveguide is chosen as 480 nm to ensure single-mode operation. The width of the hybrid waveguide is chosen as 420 nm to match the phase requirement. The gap between the two waveguides is chosen as 200 nm, which is considered the tradeoff between the insertion loss and the coupling length. The normalized optical field intensity distribution of the switch and the transmission are studied by 3D finite-difference time-domain method (Lumerical). The difference of transmission in bar state and cross state both are above 6 dB.

The device is based on an SOI platform, which is taped out from the Applied Nanotools Inc. After one EBL step, 20 nm GSSe and 10 nm silicon dioxide were deposited on the top of the silicon waveguide. The tungsten heater was completed using the second EBL step, followed by the sputtering and lift-off process. The microscope picture shows the fabricated 2 × 2 DC switch.

The off-chip fiber system measured the spectral response of the device. For each device, the initial state of PCM is at the amorphous state because of thermal deposition. We measured the transmission of the output port. After that, a series pulse with 10 V and 1 us width was sent to the heater, which allows the heater to reach the temperature of the phase transition from aGSSe to cGSSe (600 K). The measurement results of the DC switch show that for the wavelength range 1540 to 1560 nm, the extinction ratio is about 12-13 dB, and the insertion loss is relatively low (~1 dB).

Here, we demonstrated a modulo-4 shifting-RNS system based on a photonic integrated circuit (Fig.12 (b)). The modulo-4 shifting-RNS system includes six optical switches which can be operated at bar state and cross state. The insertion loss of those switches is less than 1 dB whether they are in bar state or cross state. At 1550 nm wavelength, the switch in the bar state has a 15 dB extinction ratio, and in the cross state has an 18 dB extinction ratio. The spectrum response of this switch shows a 15dB extinction ratio that can be achieved from 1540 nm to 1570 nm, which means the switch can work as a broadband (30 nm) optical switch.

The output performance of the modulo-4 shifting RNS system was measured by using an off-chip fiber system. A laser with 1550 wavelengths is individually injected from the input port 1 to port 4, and the output power of the four ports is collected simultaneously. Six switches are set as either bar state or cross state according to the LUT. For the add 0 calculation, all of the switches are set to Bar state, and the output port index of maximum power we get is the same as the input port. For other algorithms, the desired results are measured as shown in Fig. 12. (e). By setting the threshold of insertion loss to -10 dB, the signal-to-noise ratio of this module 4 shifting RNS system can reach at least 5 dB.

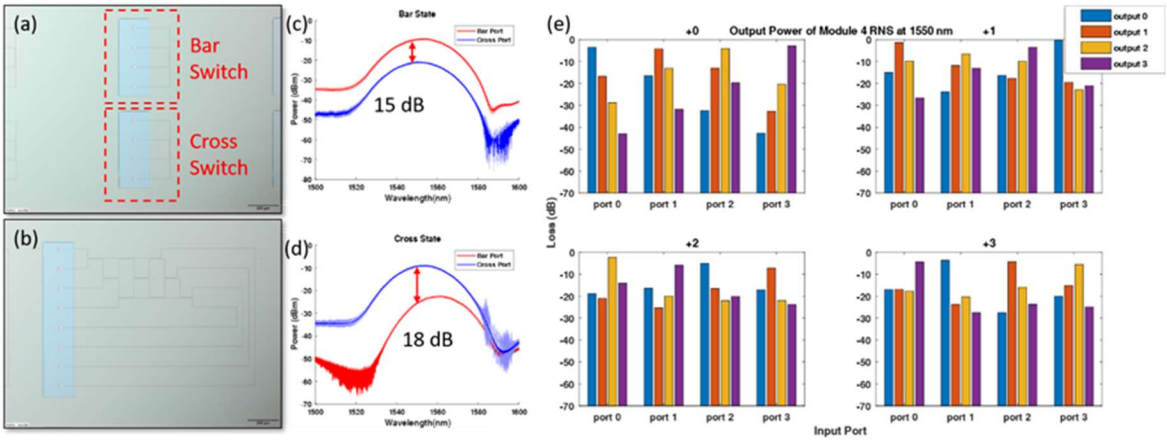


Fig 12. (a). Optical switch with bar state and cross state are used on RNS system. (b). Optical image of module 4 Shifting RNS system. (c), (d). The insertion loss of optical switch is less than 1 dB. The difference between two ports can reach 15 dB and 18 dB at Bar state and Cross state, respectively. (e). By sending light from different input ports, the maximum power is measured in the target output. The minimum difference between target output and other output is larger than 5 dB. We can easily set the threshold to distinguish 0 and 1.

4.3 Year 3

4.3.1. Optimization of an RNS DNN accelerator architecture

We presented an innovative DNN accelerator, DNNARA-E, which utilizes residue arithmetic in a hybrid optoelectrical system. The proposed work was published in ACM Journal on Emerging Technologies in Computing Systems (JETC) [13]. The proposed design employed one-hot encoding with WDM feature in RNS to accelerate the residue matrix-vector multiplication (MVM) operations while maintaining reasonable power requirements, eliminating the need for analog-to-digital converters (ADCs) and digital-to-analog converters (DACs). The DNNARA-E design superseded our preliminary work, DNNARA [8], which initially integrated RNS with photonic adders and multipliers for neural networks. However, we have identified some limitations in the previous design, including the substantial power and area consumption of optical adders in the R-MVM unit, limited support for activation functions, and a lack of comprehensive comparison with state-of-the art accelerator. Therefore, we presented DNNARA-E as an improved solution that addresses these limitations, providing a more efficient and practical design for DNN acceleration. The simulation results exhibit exceptional area and power efficiency, achieving 0.39 TOPS/mm^2 and 3.22 TOPS/W, respectively. In terms of computing capability, DNNARA-E reaches 24.91 GOPS/mm^2/W, which is at least 4.02x better than other emerging technology-based accelerators. With the same power budget as the NVIDIA GPU Tesla V100 and T4, DNNARA-E performs 80x and 72x faster on average for CNN benchmarks, respectively.

Overview. The proposed accelerator, DNNARA-E, is built based on residue arithmetic and integrated photonics, as shown in Fig. 13. The chip consists of several tiles, and each tile consists of several electro-optical residue matrix-vector multiplication units (R-MVM-E). To integrate the lasers on-chip, we utilized separated dies to place lasers with 100 μm gaps [14], avoiding the thermal stability standpoint. Recent work [15] shows the power of one on-chip laser could be up to 300 mW. Thus, one laser die is set as not exceeding the limit. Furthermore, the DNNARA-E chip comes with additional hardware units for machine learning accelerators, including activation function units, max-pooling units, LUTs, eDRAMs, R-Accumulator, and buses.

The LUTs store the switch states, which control the routing of the residue adders and multipliers. The eDRAMs store the data information of the neural network. The R-Accumulator is a residue accumulation unit that sums up the results from different tiles or R-MVM-E units. The DNNARA-E chip executes operations in the residue domain all the time to avoid the extra conversions between the binary number system and the residue number system.

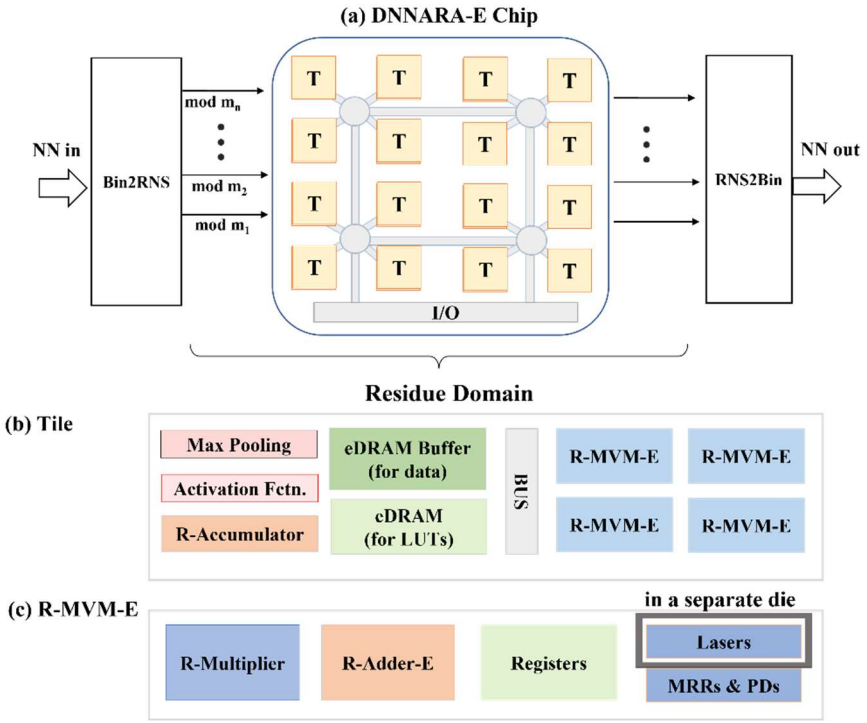


Fig. 13. Overview of the system for DNNARA-E. (a) High level design of a residue number system of a neural network with a DNNARA-E chip. Architectures of (b) a tile and (c) an electro-optical residue matrix-vector multiplication (R-MVM-E) unit. The input of the neural network will be converted to RNS and are executed in residue domain until reach the end.

Bin2RNS converters and RNS2bin converters are required to convert between the conventional binary system and the residue number system. A binary number is converted to residue number via the bin2RNS unit before storing it in the eDRAM buffer. Then all the computations are performed in the residue domain. Finally, the results will be converted back to the binary domain via the RNS2bin units once all the operations are executed.

However, conversions are not always needed inside the chips since several chips may be utilized for one neural network. Thus, we set the conversion units off-chip. Note, the conversion time/energy of different benchmarks are considered for a fair comparison while evaluating the performance in this work as an end-to-end system.

4.3.1.1. Enhanced Integrated Photonic Residue Arithmetic Computing Engine for Neural Network

Residue electrical adder. Optical residue adders and multipliers benefit from the WDM

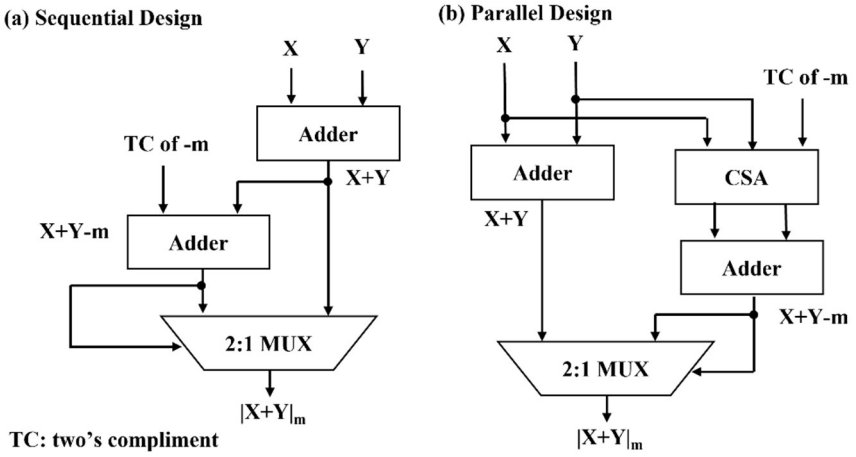


Fig. 14. Schematics of An Electric Residue Adder in (a) a sequential design and (b) a parallel design.

feature that improves performance with high parallelism. In order to achieve such benefit, the operation should have at least one common operand (e.g., +4 or * 3).

The MAC operation in a neural network fits the multiplication because the weight keeps constant in the inference stage, so that they could be operated simultaneously. However, we cannot benefit from WDM in the case of accumulation process in a neural network due to the lack of common summands. DNNARA shows that one residue optical adder could execute one addition at one time.

In this case, the large size of optical adder is found to be unjustifiable. Moreover, the electrical residue addition implementation is simple. Thus, here we propose using electrical residue adders (R-Adder-E) instead, which is smaller and more power efficient than an optical adder.

A modulo-m addition, $|X+Y|_m$, could be represented as following equation:

$$|X+Y|_m = \begin{cases} X+Y, & if\ X+Y < m \\ X+Y-m, & otherwise \end{cases}$$

Two full adders [16] and one 2:1 MUX [17] are needed to build a residue adder based on the conditions [9], as illustrated in Fig. 14. Fig. 14 (a) shows a sequential implementation where one adder is used to calculate $X+Y$, and then the result will be passed to the other adder to calculate $X+Y-m$. The MUX will choose from the two results according to the sign bit of $X+Y-m$.

A parallel implementation is shown in Fig. 14 (b) which calculates X+Y and X+Y-m in parallel. In this setting, the computation of X+Y-m is sped up using a carry save adder (CSA). Please note that computing X+Y-m will still take more time than X+Y. The speed of the circuit is dominated by the critical path of the computation of X+Y-m. The total delay time of the parallel adder will be $T_{parallel} = (n+1)T_{FA} + T_{mux}$, where n is the number

of bits, T_{FA} and T_{mux} stand for the delay of the full adder and the 2:1 multiplexer, respectively. The delay of the serial implementation is $T_{serial} = 2nT_{FA} + T_{mux}$.

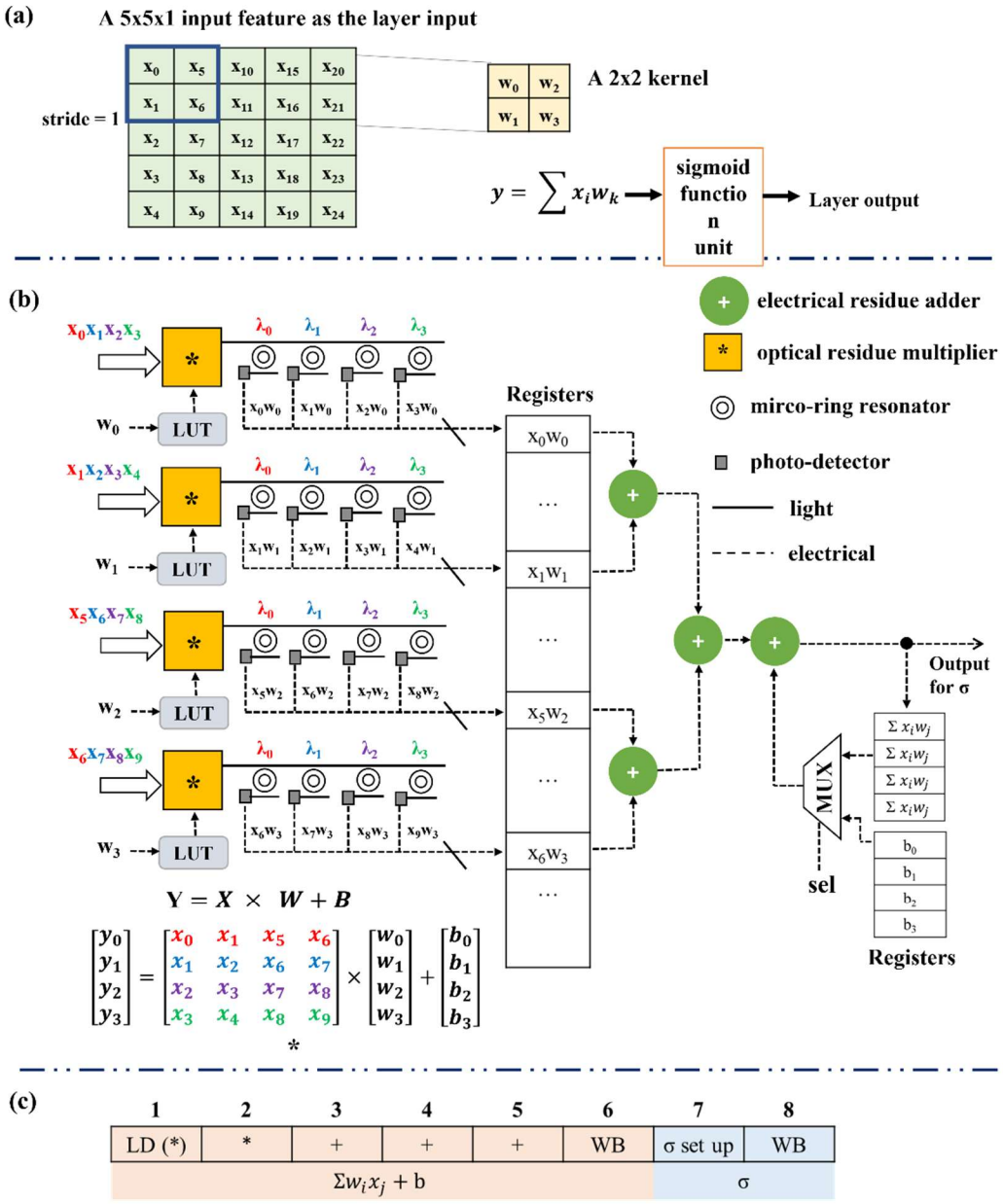


Fig. 15. (a) An example of a 5x5 input feature and a 2x2 kernel for a neural network. (b) Scheme of a Residue MVM-E Unit with Nanophotonic Residue Multipliers and Electrical Residue Adders. The convolution of (a) could be further expanded as the MVM and employed to it. The multipliers are set according to the input, and light is routed to the expected result ports. MRRs filter out the light signal with different wavelengths (λ_0, λ_1, λ_2, λ_3), and PDs detect the light. Finally, all results are accumulated by the tree-like electrical adders and sent to the next layer. (c) A pipeline design for one MAC operation and its following activation function.

Our simulation shows that the sequential design requires around 0.7x area and 0.5x power of the parallel one, while it needs 1.5x time to run a 16-bit additive operation with 32nm technology. However, both two methods could be operated in the designated one cycle (1.2 GHz). Therefore, we choose the sequential design for its smaller area and lower power requirement, but parallel design can also be used in different applications based on the accessibility to the hardware resources and limitations. R-Adder-E saves around 58% in area and 42% power consumption while operating sequentially compared to the optical design.

Residue MVM-E unit. We explored an optoelectronic residue matrix-vector multiplication unit (R-MVM-E) consisting of electrical residue adders and optical residue multipliers for the residue MAC operations in DNN.

Our prior study [8] proposed residue MVM unit to perform MAC operation in parallel using optical residue adders and multipliers with the capability of wavelength division multiplexing (WDM). However, the large area of the optical residue adder constrains the performance of the system. Many additive operations do not have any shared summands in the accumulation process of a CNN, and thus they cannot be speed up with WDM. Consequently, a *w*-wavelength residue matrix-vector multiplication unit (R-MVM) in DNNARA requires w^2 optical R-Adders. When w=5, optical residue adders consume 66% of the area and 62% of the power in an R-MVM unit.

To address the challenge, we propose a new R-MVM-E unit using electrical adders due to the simple implementation of the design. It maintains the high speed to execute residue matrix-vector multiplication operations while reducing the consumption of area and power.

The following example shows how an R-MVM-E unit with electrical adder and optical multiplier performs several MAC operations in a neural network. Assume a 5x 5 x1 input feature and a 2x2 kernel, with stride size one and zero paddings (Fig. 15 (a)). The kernel slides over the input feature and generates the output in a convolutional layer, which could be mapped as a matrix-vector multiplication and addition (MVMA). As for the first two columns of the input, the MVMA is expressed as Y = X*W+B, as shown in Fig. 15 (b). The matrix X, called a flattened matrix, is not the same as the input feature. Instead, it is rearranged according to the kernel size. The bias value is stored in vector B. Values in vector *W* are set as the electrical control signal of the optical residue multiplier, while the values in matrix X will be encoded as the optical input signals. For instance, w_0 will be set as the control signal of the first multiplier. By utilizing WDM (size = 4), the first column of X, (x_0, x_1, x_2, x_3), will be encoded as four optical signals with different wavelengths, (λ_0, λ_1, λ_2, λ_3), respectively. The MRRs at the output ports filter out the results of each operation according to the designated wavelengths. Then PDs convert the detected optical signal to an electrical one and store them in the registers.

Hence, the first multiplier operates four multiplications concurrently if the WDM size equals four. Similarly, each of the multipliers performs four multiplicative operations at the same time. Finally, several sets of tree-like electrical residue adders accumulate the results of the multiplications and generate the vector Y. The tree structure of adders helps to achieve high throughput by enabling high parallelism.

The size of matrix X is expected to be more extensive in real applications, and the hardware resources are limited by the area/power. To further explore the opportunity of a residue MAC operation for a large size vector, we split matrix-vector multiplication into smaller sizes that could be fit in a single R-MVM-E unit. Various R-MVM-E units could be used to perform sub-operations before adding the multiplied results.

At the end of each cycle, the register may store either the final result or the partial sum. A multiplexer (MUX) is used to select either the final/partial results or the bias. If w wavelengths are used in the proposed R-MVM-E design, it requires w optical R-Multipliers, w sets of MRRs and PDs, $w(w+1)$ electrical R-Adders, and (w^2+2w) memory entries. In addition to accumulating w multiplication results, one more addition is needed for either the previous y_i or the bias. Thus, each wavelength requires $(w+1)$ electrical adders to achieve the best performance, which needs $log_2 (w+1)$ stages.

Compared to the R-MVM design of DNNARA (w=5), the R-MVM-E (w=21) operates 17.64x more MAC operations at the cost of 8.84x of the area, 9% more of the power and a slightly slower executing speed (3 more cycles). The R-MVM-E performs 441 MAC operations in 8 cycles while R-MVM operates 25 ones in 5 cycles without the pipeline design. The throughput of DNNARA-E is three times higher than the one of DNNARA. Moreover, R-MVM-E can reach a higher throughput after addressing a pipeline design.

Convolutional Intra-Layer Pipeline Design. A new pipeline for the convolutional neural layer has been designed using the improved computational blocks as depicted Fig. 15 (c). Loading the switch states from a LUT consumes one cycle (around 0.5ns) according to the simulation results from CACTI 7.0 [10]. After that, the optical residue multipliers are set to corresponding states, with switch setting time as low as 5.1 ps [1]. Then the optical signal is injected and traverses the optical path in less than 1 ps [4]. An MRR resonates light at around 25GHz [18], and a PD detects light at 35~110 GHz [19]. By setting the lower bound of the 35GHz, the overall response time of the MRR and PD is around 70 ps, which is much shorter than the designated one cycle. One cycle is divided into two sub-cycles in order to guarantee that the optical switches are being set prior to light injection. The first half cycle is designed for the switch setting, while the second one is used to inject and detect the light.

The multiplicative operation loads the corresponding data and sets the states of the switches in cycle 1. Then the data flows through the optical R-Multiplier in cycle 2. After that, the data is stored in the registers for the following accumulation operations. To achieve the best performance, we utilized four electrical residue adders to accumulate the four multiplication results and the bias, consuming three cycles (cycle 3~5). Finally, one additional cycle is needed to store the results from the accumulation operations for future use. Furthermore, taking advantage of the highly parallel architecture, sixteen MAC operations could be done in six cycles simultaneously.

The sigmoid function is chosen as the activation function due to its rapid speed in the residue domain, i.e., consuming two cycles. The first cycle (Cycle 7) will be similar to cycle 2, in which optical signals are injected into the activation blocks and are routed to the desired locations, and ultimately are decoded. Then, the results are written back to

registers in cycle 8. Thus, the computations inside a convolutional neuron, including the dot-product operations with bias addition, as well as the sigmoid activation function, could be obtained in 8 cycles, which can be ideally pipelined.

4.3.1.2 Evaluation

Design Space Exploration. Here, we introduced *computational capability* as the evaluation metric, represented as the number of 16-bit fixed-point operations per second per mm^2 per Watt (GOPS/(mm^2 ▪ Watt)) that is being processed. Computational capability depends on the hardware resources, including the number of wavelengths (W), R-MVM-E in a tile (R), tiles per chip (T), appraising the computing throughput, area, and energy. To derive the maximum computational capability, we tested all parameters mentioned above at the peak performance and kept all the units working concurrently to calculate the maximum computational capability.

The simulation result shows that the optimal computational capability of one DNNARA-E chip is 24.91 GOPS/(mm^2 ▪ Watt), *with 21 wavelengths, 12 R-MVM-Es in a tile, and 8 tiles in a chip* as the configuration (Fig. 16), which is 1.98x higher than DNNARA.

The peak of computational capability varies based on different hardware configurations. In general, it increases as more hardware sources are available, but drops after some point. As depicted in Fig. 16, if (T, W) = (8, 21), the computational capability increases when R<16 and then drops. Computational capability only benefits from a limited number of R-MVM-E units as the power consumption and area increase when introducing more R-MVM-E units. The increasing proportion of the computational capability does not match the increment of WDM size due to the overhead of the mandatory MRRs and PDs for filtering results.

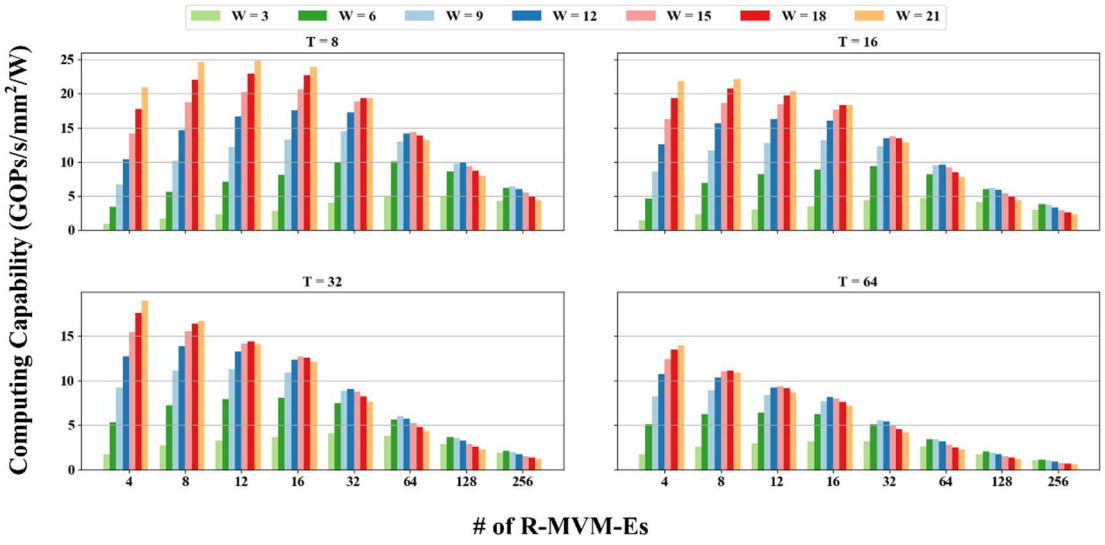


Fig. 16. Computational Capability with Different Configurations of DNNARA Design. *W* represents the number of wavelengths, *R* stands for the number of R-MVM-E units in a tile, and *T* represents the number of tiles on a chip.

Performance on real benchmarks. Fig. 17 shows the time and energy consumption of

DNNARA-E and DNNARA for the tested CNN benchmarks (LeNet5 [20], VGGs [21], DeepFace [22], ResNets [23]). Both of them are normalized to the performance of NVIDIA Tesla V100 or T4 as the baseline. The speedup is calculated by *Execution time of GPU/Execution time of DNNARA or DNNARA-E*.

In general, faster speed and higher energy consumption are expected when more chips are utilized. However, the improvement does not follow the increment of chips due to the communication overhead between chips. In most cases, DNNARA-E performs better than DNNARA in terms of time and energy because more MAC operations can be executed concurrently with R-MVM-E. However, for a small network like LeNet5, the power consumption increases more than speed improvement. Compared to Tesla T4 GPUs, DNNARA-E consumes more energy when more than 32 chips are used because too many hardware units are introduced but stay idle when they are operating.

NVIDIA reports that the power consumption of a Tesla V100 is 250 Watt. We expect that 16 DNNARA-E chips could be operated under the same power budget, and they could run 80 times faster on average. Similarly, we expect that 4 DNNARA-E chips could be run with the same power budget as a Tesla T4 GPU, which is 70W, and they could run 72x faster.

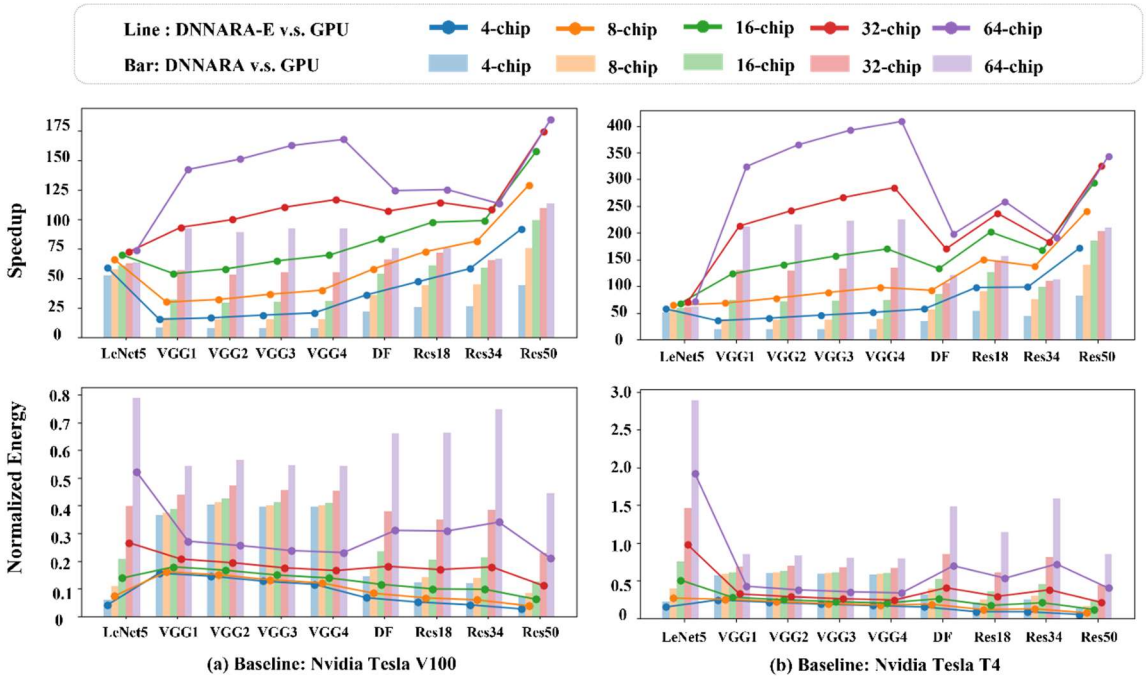


Fig. 17. The Speedup (top) and Normalized Energy Consumption (bottom) of DNNARA (bar) and DNNARA-E (line) to Run CNN Benchmarks when normalized to (a) Nvidia Tesla V100 and (b) Nvidia Tesla T4.

Overhead of Bin2RNS and RNS2Bin Conversions. DNNARA-E is built based on the residue number systems, and thus the conversion overhead should be further considered. Here, we report each benchmark's time/energy breakdown in Table 2. The number of Bin2RNS and RNS2Bin units was set by the following equations:

$$\#\ of\ Bin2RNS\ units = C \times T \times w$$
$$\#\ of\ RNS2Bin\ units = C \times T$$

where C represents the number of chips, T stands for the number of tiles on a chip, and w represents the number of wavelengths in the system. Ordinarily, both the time and energy breakdown become smaller when the network size increases because only one conversion is needed for one network. Thus, longer convolutions benefit more from the one-time conversion process. The fraction of the conversions is also related to the network input/output size. On average, the conversions take less than 2.34% of total execution time and 1.51% of total energy consumption of the neural networks. In most cases, the execution time and energy consumption of an RNS2Bin conversion are less than those of a Bin2RNS conversion. The RNS2Bin conversion process is more complicated than Bin2RNS, but fewer numbers are needed. For example, the input size of VGG1 is 150,528 while the output size is 1000. Thus, the bin2RNS conversation consumes 6.7x more time and 6.5x more energy than the RNS2bin conversion in VGG1.

Table 2. Execution Time (T) and Energy Consumption (E) Breakdown of Bin2RNS and RNS2Bin on CNN Benchmarks Using One DNNARA-E Chip. All units are in percentage

Benchmark	LeNet5	VGG1	VGG2	VGG3	VGG4	DeepFace	ResNet18	ResNet34	ReNet50
Bin2RNS(T)	1.932	0.448	0.447	0.437	0.426	1.872	3.1256	2.867	2.937
Bin2RNS(E)	1.649	0.487	0.486	0.474	0.462	0.211	3.431	3.116	3.193
RNS2bin(T)	2.657	0.067	0.067	0.065	0.064	2.320	0.471	0.428	0.439
RNS2bin(E)	0.016	0.003	0.003	0.003	0.003	0.012	0.022	0.020	0.021

4.3.2 An RNS structured network architecture for the manycore system

We further propose two architectures of residue computing nodes in the manycore system, addressing collective operations. Collective operations are an integral part of parallel computing paradigms and involve all the nodes in the parallel program. Due to the participation of all nodes, reduction operations (such as addition, multiplication, min, and max operation) are expensive. Laser controlled (Fig.18(a)) and micro-ring resonator (MRR) controlled (Fig.18(b)) design are proposed. Both of them utilize the MRR to filter out the light with different wavelengths and photo-detector at the end of the system to measure the photon current.

Multicast operation in a communication network allows a message to be sent to a specific group of processors simultaneously. For instance, P0 intends to send a message to a specific group of nodes, P1, P2, and P3. Here, we show how these two designs will process this example.

Laser controlled collective chip (Fig. 18(a)). P0 sends a command to the control unit indicating that it wants to send a message to a specific group of nodes (❶). The control unit then sets the lasers' wavelength based on the command received (❷). In the case of multicast, the lasers' wavelengths are set to correspond to the specific group of nodes that the message will be sent to. In contrast, for unicast, the laser wavelength is set to correspond to the specific destination node that the message will be sent to. After the lasers' wavelengths have been set, the light is injected into the system, and each processor's input is imprinted onto a unique wavelength (❸). The wavelengths are integrated into a

single waveguide using an arrayed waveguide grating (AWG). The micro-ring resonator filters out the light based on the designated wavelength, allowing only the specific group of nodes to receive the message. Finally, the photon-detector (PD) measures the optical current and sends it back to the corresponding processor (❹). The processors that are not part of the designated group will not receive the message because their wavelengths are not allowed to pass through the micro-ring resonator filter.

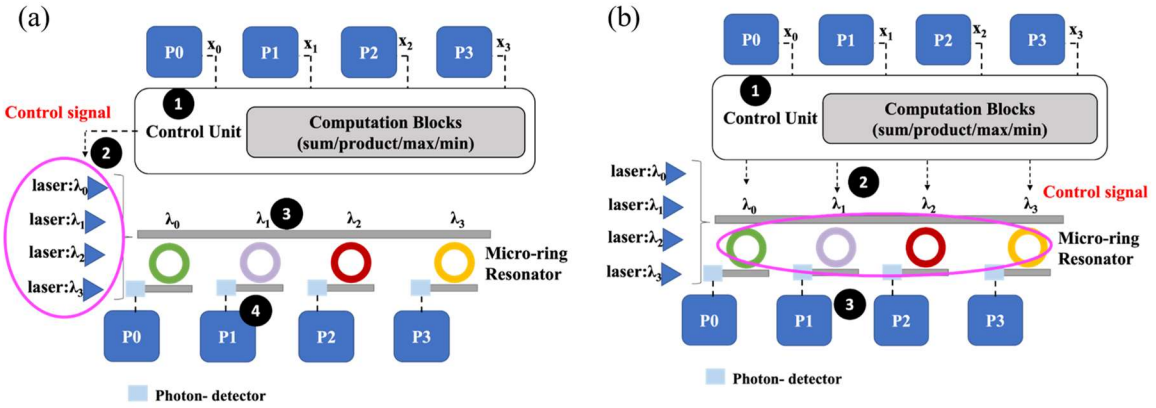


Fig. 18. Example scheme of (a) laser-controlled and (b) Micro-Ring Resonator (MRR) controlled four nodes collective chip.

MRR controlled collective chip (Fig. 18(b)). P0 transmits a directive to the control unit to convey a message to a specific set of nodes (❶). Subsequently, the control unit operates the designated MRR by administering an appropriate voltage, thereby enabling the light goes through or drops to the MRRs based on the directive received (❷). It should be noted that the MRR is an active element, which implies that the application of distinct voltage levels can alter the direction of light. The input light travels through or drops to the ring, which induces a change in light intensity for the two states, necessitating the use of ADC/DACs. In this scheme, lasers remain powered-on, and one high-powered laser per wavelength delivers optical power throughout the chip. Following the configuration of the MRR, the purple light (λ_1) descends (❸). Finally, the PD identifies the light intensity, and P1 deciphers it as the data information.

Although both architectures are capable of carrying out collective operations, their performances are not equivalent. The scheme that is governed by laser control is encumbered by lengthy configuration times for each laser, while the MRR-controlled scheme is considerably faster. Due to technological limitations, multiple rings are required to accomplish multi-bit data transmission, which necessitates a larger area and higher laser power. Our simulation results indicate that employing eight pulse amplitude modulation 4-level (PAM-4) MRRs can provide a bandwidth of up to 41GB/s, which is 14.6 times faster than that of the Blue Gene supercomputer.

4.3.3 Demonstration of an Active Silicon-based RNS CNN Chip

We designed, fabricated and measured an R-MAC (residue multiply-accumulate) chip, which is a WDM enabled silicon-based RNS matrix-vector multiplication (R-MVM) unit

at the chip level with off-chip lasers and photo-detectors. The broadband PIC design demonstrated the functionality of a 4-bit R-MVM unit, which worked as expected in 1536-1551nm range with a high signal-to-noise ratio (>7dB).

4.3.3.1 Chip Design

Fig. 19 (a) illustrates the architecture of the chip design for the active RNS MAC unit where an off-chip laser serves as the light source. A set of multiplexers (MUX) are positioned right after the laser to select the input port for the residue multipliers. The detailed design of the residue multiplier and residue adder are shown in Fig. 19(c)-(e). Here, we utilize the a heater-based Mach-Zehnder interferometer (MZI) 2x2 switch. By heating up the switch, light either goes through or goes across the switch (shown as Fig. 19(c)). The residue adder is built based on arbitrary-size Benes network (AS-Benes), which recursively decomposes the network into smaller ones (Fig. 19 (d)). The residue multiplier separates into two parts: i) one of the operand is 0, all the light is routed to output port 0; and ii) the rest cases could be considered as an (N-1)x(N-1) AS-Benes Network (Fig. 19(e)). In this design, light is transmitted through the residue multiplier and residue adder without any photonic-electric conversion until it reaches the end of the network. A set of off-chip photo-detectors (PD)

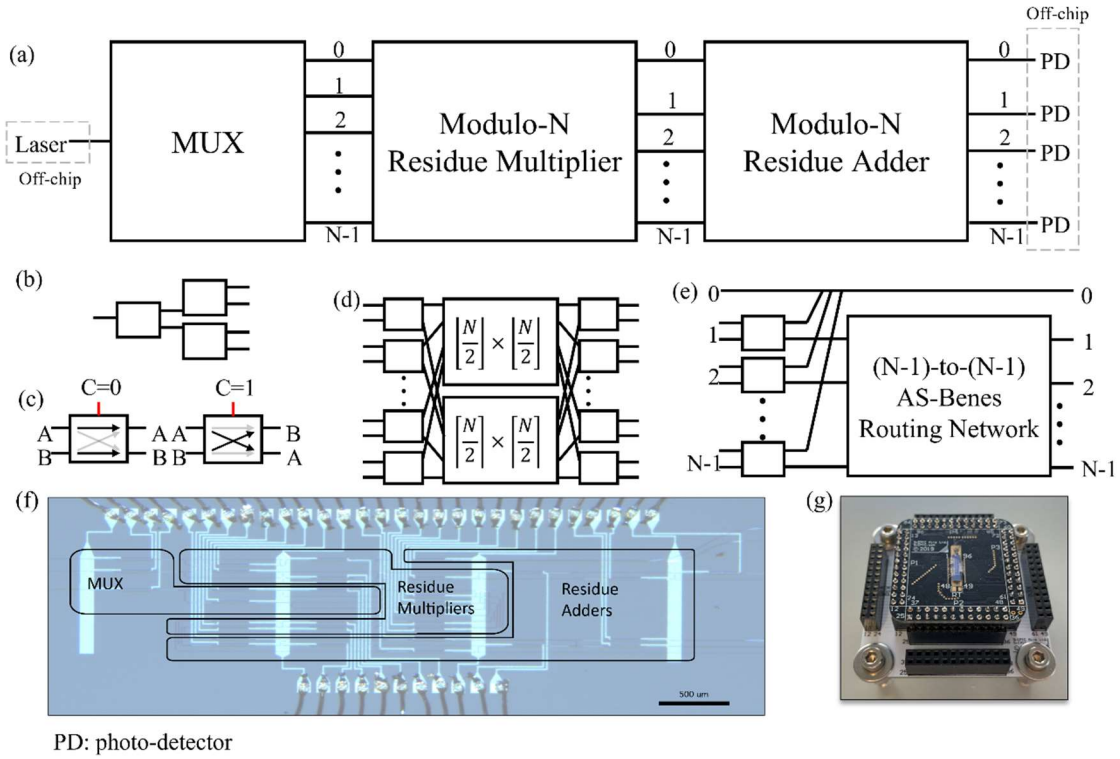


Fig 19. (a) Architectural design of the fabricated chip. (b) Example schematic of the multiplexer (MUX). (c) Two states of the 2x2 switch: *bar* state (C=0) and *cross* state (C=1). (d) Schematic of an arbitrary size Benes (AS-Benes) network. (e) Schematic of a residue modulo-N multiplier using AS-Benes architecture. (f) Microscope picture of silicon-based residue multiply-accumulation (R-MAC) chip, which includes MUX modules, residue multipliers and residue adders. (g) Optical image of electrical packaged R-MAC chip.

verifies the light signal at each output port. The design employs one-hot encoding, thereby eliminating the need for the use of ADCs/DACs. The primary objective is to determine the threshold for each output port and ensure that the signal-to-noise ratio is high enough to distinguish the position of the output light.

4.3.3.2 System Setup

To measure the performance of our designed R-MAC chip, we implemented a measurement setup as illustrated in Fig. 20. To achieve maximum coupling efficiency, we utilized polarization control (PC) to maintain the TE polarization of the input light from the off-chip laser. A center control system (a desktop in our setup) transferred the control signal of each switch to the voltage controller, which sets the state of each 2x2 switch of the MUX, residue multipliers and residue adders, according to the operands of the MAC operations. The design under test (DUT), which is the proposed active 4-bit residue MAC chip (the design under test, or DUT) then routed the light accordingly. At the end of the system, off-chip photo-detectors (PD) were placed at each output port to detect the position of the output light. All the results were transmitted to the desktop, and the light signal was decoded based on the position to deliver the results of the MAC operations.

4.3.3.3 Experimental Results

To evaluate the performance of the proposed R-MAC chip, which has been fabricated by Applied Nanotools Inc. (ANT), various tests were conducted. These included the measurement of the performance of a single switch under different states, the establishment of the threshold for each output port, testing the functionality of each possible operation, as well as verification of the WDM feature.

Performance of a single switch. The fundamental components of the R-MAC chip are comprised of the thermal-optical based 2x2 switch. As a result, it is imperative to ensure the individual switch operates as expected. To achieve this, various heater power levels were applied to the switch for both the bar state and cross state. The measurement results are displayed in Fig. 21 (a). Analysis of the measurement data revealed that at a power level of 10 mW, the insertion loss for the cross state was approximately -2 dB, while that for the bar state was -25 dB. This yielded a high signal-to-noise ratio of more than 20 dB

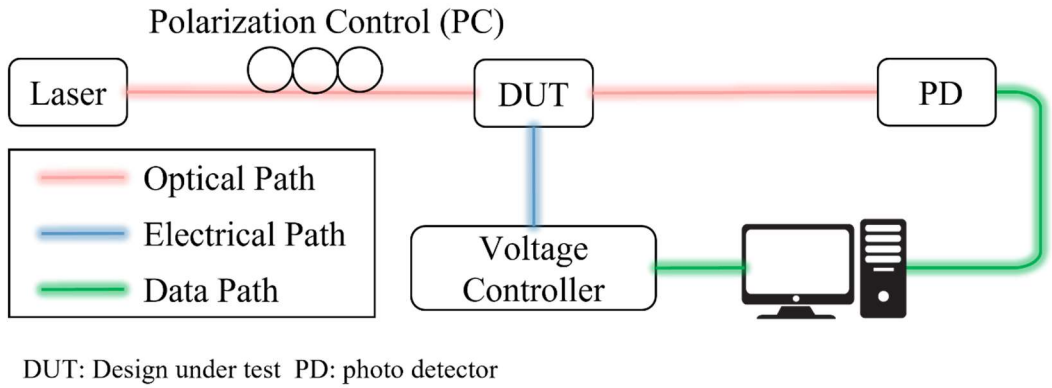


Fig. 20. The schematic of measurement setup for the fabricated chip.

range. To optimize power efficiency, the control voltage was set to 10 mW for the bar state, and 0 for the cross state.

Normalization of the R-MAC chip. The R-MAC chip is designed to operate on a 4-bit systemin. Thus, modulo-2, modulo-3, and modulo-5 systems are selected. In order to ensure that the correct threshold is set, which is required by the one-hot encoding design, the power at each output port was measured. The normalized results of these measurements are shown in Fig.21 (b). At the worst case (modulo-2 system), the normalized output power exhibited a difference of over 10dB from port to port, which is a sufficient range for determining the output value. To achieve this, a program was developed based on the

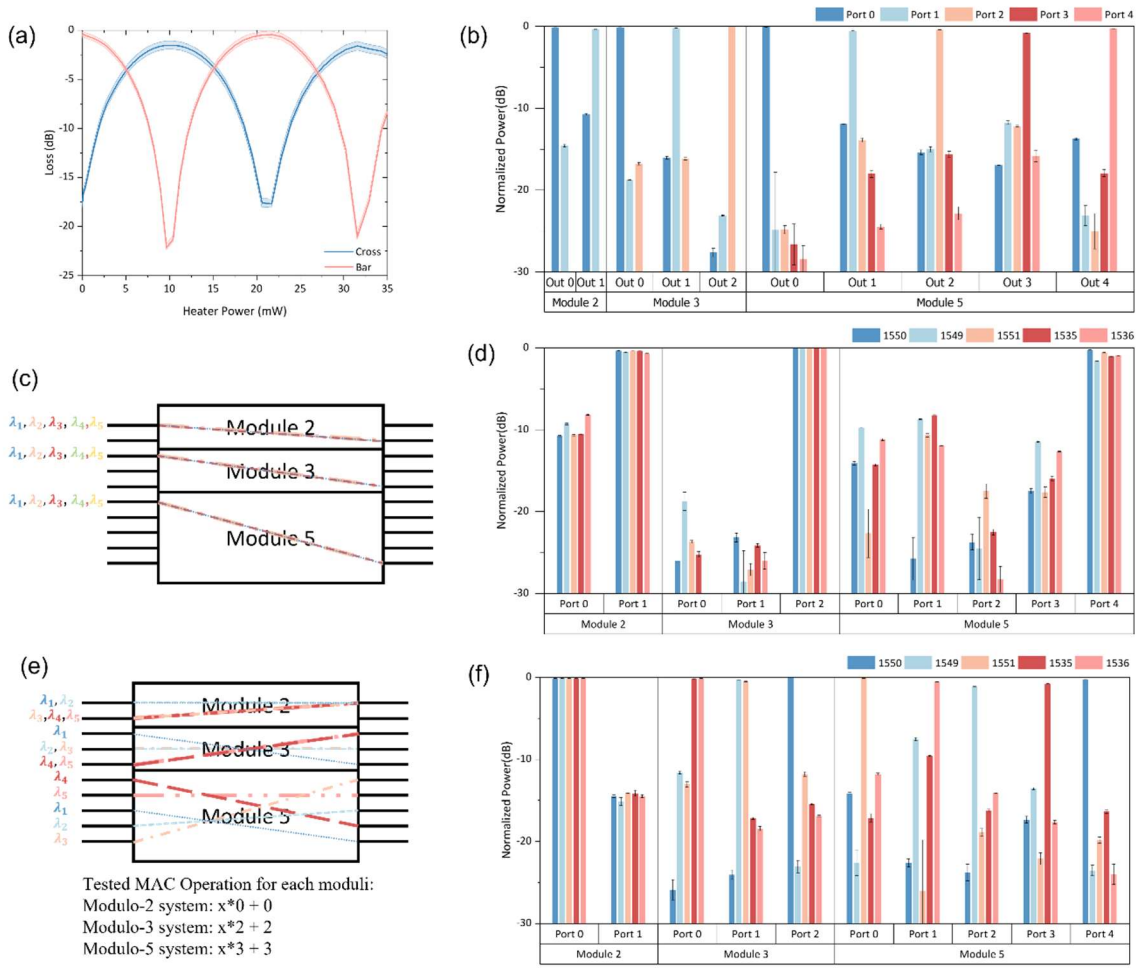


Fig. 21. (a). Performance of a single switch. The insertion loss of optical switch is less than 1 dB. The difference between two ports can reach 15 dB and 18 dB at bar state and cross state, respectively. (b). Results of function verification of RNS CNN chip. By sending light from different input ports, the maximum power is measured in the target output. The minimum difference between target output and other output is larger than 5 dB. We can easily set the threshold to distinguish 0 and 1. (c), (d). Results of longest path test under the selected wavelength. (e), (f). Results of WDM based parallel computing under the same RNS operand.

measurement data and run on the desktop. The program utilized the current derived from the photodetector to distinguish the output as either 0 or 1.

Functionality of MAC operation and WDM feature tests. To ensure the full functionality of the chip, a series of tests were conducted on each possible operation. Furthermore, the WDM feature was tested with selected wavelengths, such as1535nm, 1536nm, 1549nm, 1550nm, and 1551nm. Fig.21 (c) illustrates the setup path for each modulus when different colored light is introduced at input port 0, and Fig.21 (d) displays the corresponding results at each output port with varying wavelengths. As expected, all input lights were successfully routed to the intended output port. At the worst case (modulo-2 at 1536nm), the power difference between output port 0 and output port 1 was 8 dB. Taking into account the potential accumulation of loss nonuniformity between the bar state and cross state of the 2x2 switch, the highest power loss path for the switches was determined and presented in Fig.21 (e). All residue adders and multipliers in the same modulo system were set for the same operand. Additionally, different wavelengths were introduced into different input ports for functional testing, and measurement results are shown in Fig.21 (f). In all cases, the light was successfully routed to the desired output with a high signal-to-noise ratio. The worst-case scenario occurred in the modulo-5 system at 1549 nm, where $|3*3+3|_5 = 2$, indicating that the light was injected at input port 3 with the residue multiplier set as "*3" and the residue adder set as "+3". In this scenario, output port 2 should have received enough light to be identified as "1", while the remaining output ports should have been determined as "0". The smallest difference between the output ports was 7 dB, between output port 2 and output port 1, which was sufficient to establish the threshold.

References

[1] Sun, S., Narayana, V.K., Sarpkaya, I., Crandall, J., Soref, R.A., Dalir, H., El-Ghazawi, T. and Sorger, V.J., 2017. Hybrid photonic-plasmonic nonblocking broadband 5×5 router for optical networks. *IEEE Photonics Journal,* 10(2), pp.1-12.

[2] Tai, A., Cindrich, I., Fienup, J.R. and Aleksoff, C.C., 1979. Optical residue arithmetic computer with programmable computation modules. *Applied optics*, *18*(16), pp.2812-2823.

[3] Peng, J., Sun, S., Narayana, V.K., Sorger, V.J. and El-Ghazawi, T., 2018. Residue number system arithmetic based on integrated nanophotonics. *Optics letters*, *43*(9), pp.2026-2029.

[4] Peng, J., Alkabani, Y., Sun, S., Sorger, V.J. and El-Ghazawi, T., 2019, November. Integrated Photonics Architectures for Residue Number System Computations. In *2019 IEEE International Conference on Rebooting Computing (ICRC)* (pp. 1-9). IEEE.

[5] Peng, J., Sun, S., Narayana, V.K., El-Ghazawi, T. and Sorger, V.J., 2019, August. Silicon Photonic Enabled Residue Number System Adder and Multiplier. In *2019 IEEE Research and Applications of Photonics in Defense Conference (RAPID)* (pp. 1-2). IEEE.

[6] Beneš, V.E., 1964. Permutation groups, complexes, and rearrangeable connecting networks. *Bell System Technical Journal*, *43*(4), pp.1619-1640.

[7] Chang, C. and Melhem, R., 1997. Arbitrary size benes networks. *Parallel Processing Letters*, *7*(03), pp.279-284.

[8] Peng, J., Alkabani, Y., Sun, S., Sorger, V.J. and El-Ghazawi, T., 2020, August. DNNARA: A Deep Neural Network Accelerator using Residue Arithmetic and Integrated Photonics. In *49th International Conference on Parallel Processing-ICPP* (pp. 1-11).

[9] Omondi, A.R. and Premkumar, A.B., 2007. Residue number systems: theory and implementation (Vol. 2). World Scientific.

[10] Muralimanohar, N., Balasubramonian, R. and Jouppi, N., 2007, December. Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007) (pp. 3-14). IEEE.

[11] Li, J., Martinez, J.F. and Huang, M.C., 2004, February. The thrifty barrier: Energy-aware synchronization in shared-memory multiprocessors. In 10th International Symposium on High Performance Computer Architecture (HPCA'04) (pp. 14-23). IEEE.

[12] Anbar, A., Serres, O. and El-Ghazawi, T., 2011, December. Reflex Barrier: A Scalable Network-Based Synchronization Barrier. In 2011 IEEE 17th International Conference on Parallel and Distributed Systems (pp. 204-211). IEEE.

[13] Peng, J., Alkabani, Y., Puri, K., Ma, X., Sorger, V. and El-Ghazawi, T., 2022. A Deep Neural Network Accelerator using Residue Arithmetic in a Hybrid Optoelectronic System. ACM Journal on Emerging Technologies in Computing Systems (JETC), 18(4), pp.1-26.

[14] Blaicher, M., Billah, M.R., Kemal, J., Hoose, T., Marin-Palomo, P., Hofmann, A., Kutuvantavida, Y., Kieninger, C., Dietrich, P.I., Lauermann, M. and Wolf, S., 2020. Hybrid multi-chip assembly of optical communication engines by in situ 3D nano-lithography. Light: Science & Applications, 9(1), p.71.

[15] Kharas, D., Plant, J.J., Loh, W., Swint, R.B., Bramhavar, S., Heidelberger, C., Yegnanarayanan, S. and Juodawlkis, P.W., 2020. High-power (> 300 mW) on-chip laser with passively aligned silicon-nitride waveguide DBR cavity. IEEE Photonics Journal, 12(6), pp.1-12.

[16] Ulloa, G., Lucena, V. and Meinhardt, C., 2017, December. Comparing 32nm full adder TMR and DTMR architectures. In 2017 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS) (pp. 294-297).

[17] Ghafouri, T. and Manavizadeh, N., 2021. Design and simulation of high-performance 2: 1 multiplexer based on side-contacted FED. Ain Shams Engineering Journal, 12(1), pp.709-716.

[18] Jayatilleka, H., Caverley, M., Jaeger, N.A., Shekhar, S. and Chrostowski, L., 2015, April. Crosstalk limitations of microring-resonator based WDM demultiplexers on SOI. In 2015 IEEE Optical Interconnects Conference (OI) (pp. 48-49). IEEE.

[19] Salamat, S., Imani, M., Gupta, S. and Rosing, T., 2018, November. Rnsnet: In-memory neural network acceleration using residue number system. In 2018 IEEE International Conference on Rebooting Computing (ICRC) (pp. 1-12). IEEE.

[20] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[21] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[22] Taigman, Y., Yang, M., Ranzato, M.A. and Wolf, L., 2014. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern

recognition (pp. 1701-1708).

[23] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).