NeOCNN: NTT-enabled Optical Convolution Neural Network Accelerator

Xianbin Li
Hong Kong University of Science and Technology
Hong Kong, China
xianbin.li@connect.ust.hk

Yinyi Liu
Hong Kong University of Science and Technology
Hong Kong, China
yinyi.liu@connect.ust.hk

Fan Jiang
Hong Kong University of Science and Technology
Hong Kong, China
fjiangad@connect.ust.hk

Chengeng Li
Hong Kong University of Science and Technology
Hong Kong, China
clicu@connect.ust.hk

Yuxiang Fu
Hong Kong University of Science and Technology
Hong Kong, China
yuxiang.fu@connect.ust.hk

Wei Zhang
Hong Kong University of Science and Technology
Hong Kong, China
eeweiz@ust.hk

Jiang Xu*
Microelectronics Thrust, Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
jiang.xu@ust.hk

ABSTRACT

In the realm of neural network computation, optical neural network accelerators (ONNs) have emerged as a promising solution, leveraging the inherent speed and parallelism of optical systems. Despite their potential, current ONN designs often fall short due to inefficient data movement and reliance on traditional electronics-based dataflows.

Herein, we introduce a pioneering approach to ONN implementation that incorporates the Number Theoretical Transform (NTT), known for its effective divide-and-conquer strategy, non-floating-point operations, and multi-level parallelism. This integration significantly reduces data mapping costs for Convolutional Neural Networks (CNNs), making it particularly well-suited for optical convolution neural networks (OCNNs) with considerable performance and efficiency escalation.

For the first time, our proposed methodology realizes the benefits of NTT in ONN. We employ an innovatively designed optical butterfly structure that facilitates real-time NTT computation while occupying a minimal footprint. The resulting system, termed NEOCNN, showcases a remarkable throughput capability of up to 61 Tera Operations per Second (TOPs) and demonstrates a power efficiency of 9.6 TOPs/Watt, all while maintaining reliable inference accuracy.

This work not only represents a stride toward more efficient ONNs but also sets a precedent for future research in combining Number Theoretical Transforms with optical computing paradigms.

CCS CONCEPTS

• **Hardware** → **Emerging optical and photonic technologies**; **Emerging architectures**; **Emerging tools and methodologies**; • **General and reference** → *Design*; *Evaluation*.

KEYWORDS

Optical Neural Network (ONN), Number Theoretical Transform (NTT), Convolution Neural Network (CNN), hardware acceleration

ACM Reference Format:
Xianbin Li, Yinyi Liu, Fan Jiang, Chengeng Li, Yuxiang Fu, Wei Zhang, and Jiang Xu. 2024. NeOCNN: NTT-enabled Optical Convolution Neural Network Accelerator. In *Proceedings of the 38th ACM International Conference on Supercomputing (ICS '24), June 04–07, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3650200.3656609

1 INTRODUCTION

Neural networks (NNs) have revolutionized a multitude of fields, enabling advancements in autonomous driving, style transferring, medical diagnosis, and more, through their sophisticated pattern recognition and decision-making capabilities. Pioneering applications such as ChatGPT[2] and DALL-E[1] underscore the transformative impact of these technologies. However, as traditional computational scaling laws wane, e.g. the breakdown of Dennard scaling[12] and the stagnation of Moore's law[32], the substantial computational demands of neural network training and inference pose a challenge for real-time deployment of NNs.

*Corresponding author

ICS '24, June 04–07, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-0610-3/24/06
https://doi.org/10.1145/3650200.3656609

To address this, dedicated hardware accelerators like FPGAs [8, 41], ASICs[37], and Compute-in-Memory (CIM) implementations [23, 35] have been developed to optimize performance and reduce power consumption. Among these, silicon photonics stands out, enabling abundant computing resources with an ultra-high working frequency compared to traditional electronic ones[53]. The unique high bandwidth and low latency, together with its compatibility with CMOS processes, make it extremely advantageous for integrated electro-optical co-design and fabrication[6, 42]. Recent advances in photonics, including ones that are designed for the acceleration of CNNs[7], SNNs[16] and even transformer-based models[3], have demonstrated the potential for terabyte (TB)-[46] or even petabyte (PB)-[10] level computational performance, indicating a promising path forward for meeting the increasing computational demands of neural networks.

Despite these advancements, optical neural networks (ONNs), especially optical convolutional neural networks (OCNNs), still face inefficiencies due to the data mapping tailored for electronic processors. The high operating frequency of optical devices enables outstanding inference speeds but also introduces significant data movement and electro-optical conversion bandwidth overlap, leading to stress on memory access and a failure to reach the performance upper bound when integrated into a computer architecture.

Conventional dataflow approaches like input-stationary and weight-stationary are intuitive but lack optimization opportunities in an optical system. While row-stationary dataflow shows promise in CiM-based accelerators[11], it remains less effective for optical systems.

A promising approach to overcoming these challenges in OCNNs is the use of Fast Fourier Transform (FFT)-based convolution acceleration. Compared to the traditional spatial mapping scheme, the overall memory accesses are reduced from $O(N^2)$ to $O(NlogN)$ within a single convolution operation, which greatly alleviates the heavy memory burden. However, the phase modulation of the Mach-Zehnder Interferometer (MZI) devices is greatly impacted by the thermal crosstalk and may lead to a significant accuracy degradation[29]. Moreover, the huge footprint of MZI devices confines it from higher integration.

Number Theoretical Transform (NTT), which involves only integer operations instead of the floating-point complex computations in FFT, offers a compelling alternative. Extensively adopted in the expedition of polynomial multiplication in post-quantum-cryptography (PQC) schemes, especially fully homomorphic encryption (FHE) schemes[51], NTT has demonstrated its effectiveness in the acceleration of the convolution process of CNN[18, 45] as well. Eliminating the need for complex-real number conversions, the simplicity of NTT reduces overall transformation and memory overhead. Furthermore, the intrinsic multiple-parallelism and its simplified data traffic and mapping make NTT-based convolution acceleration remarkably suitable for optical implementation.

In this paper, we present an NTT-enabled optical convolution neural network accelerator, NEOCNN, and the contribution of the paper is outlined as follows:

- *For the first time*, we incorporate NTT into the acceleration of convolution in an optical system for reduced computational resource consumption and optimized dataflow, which consequently improves system flexibility, scalability, and accuracy.
- We propose a photonic butterfly structure and adopt it with an NTT mesh to achieve on-the-fly NTT transformation. With the optimization of the inverse design technique, the butterfly, comprised of nanophotonic devices, substantially reduces the hardware cost and footprint of NEOCNN.
- We carry out a detailed analysis of the potential crosstalk and loss factors, and verify the NEOCNN design with outstanding inference accuracy results.
- NEOCNN reaches up to 61 Tera Operations per Second (TOPs) in throughput and 9.6 TOPS/W in power efficiency for VGG-16 inference, which surpasses both SOTA electrical accelerator and ONN.

In the remaining sections, the paper is organized as follows. **Section II** provides detailed background on both Optical neural network (ONN) acceleration and the application of number theoretic transform (NTT). **Section III** thoroughly illustrates the system design of NEOCNN. **Section IV** presents a comprehensive analysis of potential crosstalk and loss factors. Finally, **Section V** concludes NEOCNN's superior performance.

2 PRELIMINARY

2.1 Optical Neural Network Accelerators

Integrated photonics, leveraging its high bandwidth, high parallelism, and low latency, has emerged as a powerful platform for accelerating artificial intelligence applications, including CNNs[7, 46], transformer-based models[3], etc. General Matrix Multiplication accelerators are another approach to fully utilize the versatility of photonic computing cores with the ability of generalization[47].

In the realm of optical neural networks (ONNs), there are two prevalent architectures: Mach-Zehnder Interferometer (MZI)-based and MicroRing (MR)-based systems. The matrix-vector multiplication (MVM) process is explored as follows, which is a cornerstone computation in neural networks, as implemented by these two types of ONNs.

MZI-based ONN: The MZI-based ONN utilizes a device called a Mach-Zehnder Interferometer (MZI) for modulating light. An MZI is a silicon-based optical modulator with two inputs and two outputs. By adjusting the optical phase in its arms, an MZI can act as a 2×2 unitary matrix operator, which can manipulate the amplitude and phase of light passing through it. The neural network inference process, particularly the matrix multiplication operation, can be decomposed into three key steps based on the Singular Value Decomposition (SVD) and the parametrization of unitary matrices. These steps involve unitary and diagonal matrices being represented by a sequence of MZI modulators, the operation of which is depicted in Figure 1. After training, the neural network's weights are encoded into phase shifts within the MZIs for inference tasks.

MR-based ONN: As depicted in figure 2a, the MR encompasses a ring waveguide, closely coupled with an input and a drop bus waveguide. The ring waveguide resonates only when the path length of the resonator cavity is identical or close to the integer multiple of

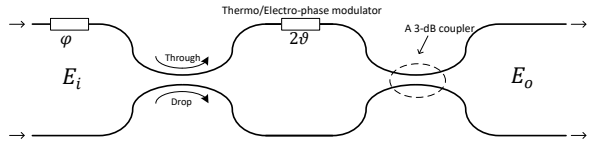
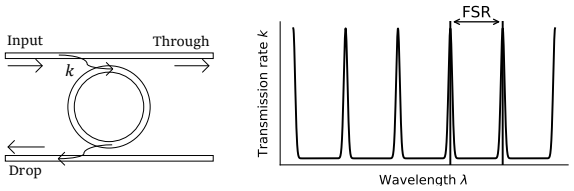
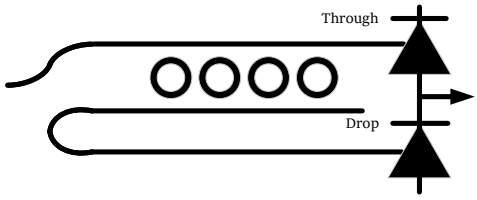


Figure 1: MZI Resonator: $E_o = j \begin{bmatrix} e^{j\varphi} sin\vartheta & cos\vartheta \\ e^{j\varphi} cos\vartheta & -sin\vartheta \end{bmatrix} E_i$

the input wavelength. This resonance happens when the circumference of the ring is an integer multiple of the wavelength of the incoming light. During resonance, light is directed from the input waveguide into the ring and then to the drop waveguide. By adjusting the resonant frequency through thermal or electrical means, we can control the intensity of light, denoted by the transmission rate k, that passes through. This intensity modulation serves as the 'weight' in MVM operations. An MVM is thus performed by modulating multiple light signals in a silicon 'weight bank', where multiple dot-products are carried out simultaneously and the differential operation of port 'Through' and 'Drop' enables sign-number computation. These vector products are then combined to execute the convolutional or MLP layers of the network.



(a) MicroRing Resonator **(b) Transmission spectrum of MR**



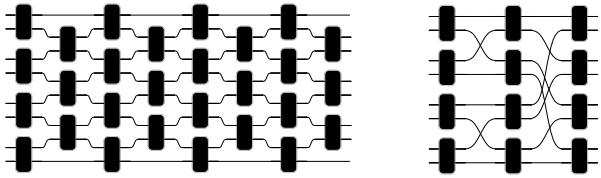
(c) Schematic of silicon weightbanks

Figure 2: Working principle of MR-based ONN

While optical neural networks (ONNs) have shown great promise in advancing artificial intelligence, they are not without their challenges. For instance, MR-based networks struggle with precision in modulation and the challenge of handling heavy weight loads. Zhang et al. [50] made significant progress by developing MR synapses capable of beyond 9-bit precision. Yet, there is still considerable overhead regarding the dithering signal as well as the intensive OE conversion. Recent work by Xu *et al.* [46] takes great advantage of the parallelisms in MR and achieves an 11.3 TOPs performance in theory.

On the other hand, MZI-based networks face issues with the complex methods required for matrix decomposition along with its inefficiency. To address this problem, Liu et al. have explored more efficient decomposition methods[28], such as sine-cosine decomposition, to reduce the size of these networks.

Additionally, alternative approaches such as FFT (Fast Fourier Transform)-based convolution[17, 36] are expected to further minimize the network footprint. The FFT-based optical convolution neural networks (OCNN) have been proposed to leverage the reduced time complexity of FFT convolution[15], and demonstrate outstanding performance and energy efficiency compared to traditional SVD-based MZI networks[40]. As illustrated in Figure 3, the FFT-based networks could reduce the number of basic computing units from $O(N^2)$ to $O(NlogN)$ during decomposition, where N represents the size of the input data. These computing units are depicted as black boxes in the figure 3. In this context, 2×2 optical modulator, conventionally MZI, is adapted to the networks for the passive optical signal transform. Whereas, in FFT-based ONNs, the thermal crosstalk that exists commonly would cause a significant precision degradation in the phase modulation of MZI modulators, which is not desirable. Moreover, the size of MZI devices remains a barrier to achieving higher levels of integration compared to MR devices.



(a) SVD-based **(b) FFT-based**

Figure 3: Footprint comparison of basic transform unit on both topologies: (a) Unitary transform; (b) FFT transform.

2.2 Number Theoretical Transform

In lattice-based cryptosystems, the Number Theoretic Transform (NTT) plays a pivotal role in accelerating polynomial multiplications, significantly reducing the time complexity of these operations. Other than FFT, which is also adopted in the acceleration of convolution computation, NTT is performed within the integer domain rather than the complex domain, intrinsically avoiding the floating point computation overhead and the according errors. Such improvements are particularly crucial in the context of Fully Homomorphic Encryption (FHE) schemes, where polynomial multiplications account for a significant portion of computational resources. For instance, in the CKKS encryption scheme, as implemented in Microsoft SEAL, an open-source cryptography library, polynomial multiplication accounts for 54.01% of the total processing time[38].

Similar to the Discrete Fourier Transform (DFT), NTT can benefit from a divide-and-conquer approach, utilizing algorithms such as Cooley-Tukey (CT) and Gentleman-Sande (GS) for efficient transformation computation. Compared to the complicated complex number arithmetic of FFT, a fast NTT algorithm could leverage the divide-and-conquer strategy to transform the computational complexity of discrete convolutions from $O(N^2)$ to $O(N \log N)$ while requiring only integer operations. This significantly reduces the computation overhead. By decomposing a problem into smaller

instances of the same problem and then combining the solutions of these instances, NTT efficiently computes the convolution of two sequences. In the context of convolution, NTT performs the transform without the detailed frequency domain information compared to the FFT, which is not required during the entire process. The efficiency gain is particularly advantageous for processing large-scale data, where direct computation methods would be prohibitively slow or even impossible.

Table 1: Analysis of convolution schemes: Transform cost, Operation cost, and memory requirement

	Winograd	**FFT**	**NTT**
Domain Transform Cost	$2(K+M-1)^2$	$3N^2 \log N$	$\frac{3}{4}N^2 \log N$
Operation Cost	$(K+M-1)^2$	$3N^2$	N^2
Memory Requirement	$2(K+M-1)^2$	$4N^2$	$2N^2$
Precision Loss	No	Yes	No

NTT outperforms other schemes in terms of operation cost and memory requirement[18], including FFT, Winograd, and Spatial, as detailed in Table 1. Assume a 2-D convolution process, where an $I \times I$ input feature map is convolved with a $K \times K$ weight kernel, producing a $M \times M$ output. The transform length of both FFT and NTT are set as $N = L + K - 1$.

To address the challenges of applying FHE in practical scenarios, researchers have proposed various NTT acceleration methods tailored to different hardware platforms. These include CPUs, GPUs, FPGAs, ASICs, and Compute-in-Memory (CIM) implementations. Nejatollahi *et al.*[33] propose the first RRAM-based NTT accelerator using in-memory bit-wise operations, while Park *et al.*[35] introduce a VMM-based RRAM NTT accelerator with a modified Montgomery reduction algorithm. However, their design only supports polynomial orders up to 1k due to the inability to reprogram RRAM arrays for larger twiddle factor matrices. Li *et al.*[23] propose MeNTT using 6T-SRAM arrays, which can operate $n/2$ butterflies in parallel but still require $log_2 n$ serial execution stages. In addition to CiM accelerators, there exist NTT implementations based on ASIC and FPGA. Banerjee *et al.*[8] propose a reconfigurable cryptographic processor, while Song *et al.*[41] construct a three-stage configurable NTT core. Zhang *et al.*[49] propose a five-stage pipeline butterfly arithmetic unit and employ a ping-pong memory access scheme on an FPGA platform. Other works[5, 51] propose customized data flow and optimized memory access strategies. However, most of these solutions must navigate a delicate balance between the flexibility and performance of the accelerators. This trade-off arises because pursuing maximum computational performance often requires platform/application-specific customization, which can reduce the generality and adaptability of the acceleration scheme.

Moreover, the multiple intrinsic parallelism of NTT could be utilized with optical implementations. Li et al. proposed to accelerate NTT with photonic implementation[24], whereas traditional MVM-based NTT is applied and the algorithmic overhead restricts it from reaching a higher performance bound when scaling up. Moreover, the reduction unit inside grapples with the inefficient general algorithm.

3 NEOCNN

3.1 NTT-based convolution and Fermat number

The Discrete Fourier Transform (DFT) is widely used across various engineering disciplines due to its ability to transform discrete signals between time and frequency domains. Similarly, The Number Theoretic Transform (NTT) extends the principles of DFT to the integer domain. The NTT replaces the complex exponential factors of the DFT, denoted as $e^{-2\pi ik/N}$, with integer values. Specifically, these integers are roots of unity within the quotient ring $\mathbb{Z}/q\mathbb{Z}$, where q is a prime number congruent to 1 modulo $2n$. Here, w is selected as the n-th primitive root of unity that satisfies the congruence $w^n \equiv 1 \pmod{q}$.

Consider a polynomial $a(X) = \sum_{i=0}^{n-1} a_i X^i$. The NTT computes an n-point transform of a as follows:

$$\tilde{a}_i = \sum_{j=0}^{n-1} a_j w^{ij} \quad \text{for} \quad 0 \le i < n \tag{1}$$

In this formula, w^{ij} are the twiddle factors—integral powers of w—and serve a role analogous to the exponential factors in the DFT.

NTT-based convolution is akin to polynomial multiplication, which is detailed in Algorithm 1.

Algorithm 1 NTT-based Convolution

Input: Vectors $a(\cdot), b(\cdot) \in \mathbb{Z}/q\mathbb{Z}$ to be convolved
Constants: Modulus q, vector length n, and primitive n-th root w
Output: Convolution result $c(\cdot)$
Function Convolution$(a(\cdot), b(\cdot), n, q, w)$**:**
  Padding(a); Padding(b);
  $A(\cdot) \leftarrow$ NTT$(a(\cdot), q, w)$;
  $B(\cdot) \leftarrow$ NTT$(b(\cdot), q, w)$;
  $C(\cdot) \leftarrow$ HadProd$(A(\cdot), B(\cdot))$;
  $c(\cdot) \leftarrow$ INTT$(C(\cdot), q, w)$;
  return $c(\cdot)$;

The concept of one-dimensional convolution extends to two dimensions as expressed in equation 2:

$$H(m,n) = \text{NTT\_2D}(h(j,k)) = \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} h(j,k) w^{mj+nk} \tag{2}$$

3.2 Fermat number theoretic transform

In Fully Homomorphic Encryption (FHE), a sufficiently large modulus q is crucial for maintaining security. The Residue Number System (RNS) is commonly employed to manage multiple smaller

moduli, thereby simplifying the computational complexity. Conversely, in NTT-based convolution, selecting a single optimal q can streamline both the transform and modular reduction processes, improving efficiency with specialized algorithms.

Particularly advantageous is the choice of a Fermat prime as the modulus q. A Fermat prime is of the form $F_n = 2^{2^n} + 1$, and for $n \leq 4$, each F_n is a prime number with desirable computational properties. For example, $F_4 = 65537$ is a prime that supports efficient NTT operations since it allows the use of primitive roots like $w = 4$ and requires only simple bit-wise operations for addition, negation, and reduction. This is because multiplication by 2^k and reduction modulo F_n can be implemented with basic bit shifts and subtractions when q is a Fermat prime, as demonstrated in Agarwal and Burrus's work on fast convolution algorithms[4].

In the specific context of CNNs, where parameters are typically represented with less than 16 bits—often quantized to 8 bits or lower—the Fermat prime $F_4 = 65537$ is an exceptional fit as the modulus. The bit width $B = 17$ required for representing F_4 aligns seamlessly with the precision of these networks, striking a delicate balance between computational efficiency and adequate precision of the networks.

3.3 Overlap-and-Add (OaA) method

In CNN architectures, the disparity between the relatively large input feature map and the much smaller kernel size can lead to computational inefficiencies when utilizing balanced NTT-based convolutions. The conventional approach, detailed in Algorithm 1, often necessitates transforming the small kernel matrix into a large-point NTT form. This scaling mismatch can result in significant performance degradation. To address this issue and integrate NTT more effectively into CNN models, the Overlap-and-Add (OaA) technique is employed[34]. This method partitions extended input sequences into smaller segments and then overlaps the convolution outcomes, as described in Algorithm 2.

Algorithm 2 NTT-Based Convolution With OaA

Input: Input Matrix $M(I \times I \times C_{\text{in}})$, Weight $W(K \times K \times C_{\text{in}} \times C_{\text{out}})$
Output: $R((I+K-1) \times (I+K-1) \times C_{\text{out}})$
for $z \leftarrow 1$ *to* C_{out} **do**
  for $k \leftarrow 1$ *to* C_{in} **do**
    $\tilde{W}(k,z) \leftarrow NTT(W(k,z))$;
  end
  for $(i,j) \leftarrow (1,1)$ *to* (I,I) **do**
    $\tilde{P}(i,j) \leftarrow 0$;
    for $k \leftarrow 1$ *to* C_{in} **do**
      $\tilde{M}(i,j,k) \leftarrow \texttt{NTT}(M(i,j,k))$;
      $\tilde{P}(i,j) \leftarrow \tilde{P}(i,j) + \texttt{HadProd}(\tilde{M}(i,j,k), \tilde{W}(k,z))$;
    end
    $P(i,j) \leftarrow \texttt{INTT}(\tilde{P}(i,j))$;
    $R(i,j,z) \leftarrow \texttt{OaA}(P(m,n)|^{i,j}_{m=i-1,n=j-1})$;
  end
end
return R;

In this method, the input feature map of size $I \times I$ is divided into smaller tiles of size $L \times L$. Each tile $\tilde{M}(i,j,k)$, along with the kernel W, is transformed into the NTT domain. The Hadamard product (denoted as Had_Prod in the pseudo-code) is an element-wise multiplication between the transformed tiles and weights. The resulting partial products $\tilde{P}(i,j)$ are then inversely transformed back to the spatial domain to obtain $P(i,j)$. These are overlapped and added together to accumulate the final output $R(i,j,z)$. The convolution is performed progressively, processing all the input tiles to produce the convolved result.

3.4 Inverse-design Nanoswitch and butterfly

Though modulators like MicroRing (MR) or Mach-Zechder Interferometer (MZI) have been widely applied in previous ONN implementations and have proven their effectiveness and efficiency, they suffer from their huge footprint of several hundreds of square micrometers. The nanophotonic devices, which function mainly with their digitized meta-structures at a sub-wavelength scale[19], demonstrate their potential in place of these modulators. These digitized meta-structured nanophotonic components have succeeded in serving as optical processors to achieve bar state, cross state, and unitary transmission[31], which is functionally similar to an MZI whereas thousands of times smaller in footprint.

One of the most significant approaches to designing nanophotonic devices is by inverse design, which inversely optimizes optical devices in terms of shape or topology according to the optical responses of the physical structure. These inverse-designed optical components have been adopted in various scenarios, including programmable nanophotonic processors, quantum photonic circuits, and photonic neural network circuits. Though a significant footprint reduction is achieved with its adoption in previous ONNs[48, 52], it is found hard to scale up due to the significant inverse-design overhead and thus only naive networks are presented. Constrained by its tiny physical size and the limited design space, previous designs[48, 52] only design networks with naive models and suffer from inaccurate modulation when the system scales up and the error accumulates. In Fig. 3b, both meshes are composed of 2×2 MZIs modulators, which could be replaced by nanophotonic devices using inverse design. As is seen in Fig.4b, the nanoswitch is similar to MZI topologically and both act as 2×2 optical modulators.

The basic butterfly modules, illustrated in Fig. 4a are responsible for (I)NTT transform in electronic accelerators. In NEOCNN, optical modulators, specifically the nanoswitches, replace them to pursue higher working frequency and lower energy consumption. In summary, the transform will be conducted passively, which means that no extra energy consumption is required except for the signal input, and rapidly, which means that the inference is processed at light speed without any delay. The nanoswitch, based on a standard silicon-on-insulator (SOI) platform, consists of two layers: a 220-nm-thick silicon layer and a silica cladding layer for protection. As depicted in Fig. 4b, the nanoswitch consists of two input/out arms and a digitized meta-structure formed by specifically distributed nanoholes. The distribution is obtained by a topology optimization algorithm as outlined in Alg. 3.

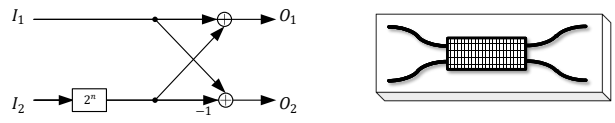


Figure 4: (a) NTT butterfly; (b) Optical inver-design nanoswitch

Algorithm 3 Adjoint Method for Inverse Design in Nanophotonics

Input: Initial design $\epsilon^{(0)}$, objective function F, constraint equations M, convergence tolerance δ, maximum iterations $N_{\max}$
Output: Optimized design ϵ^*
Procedure *AdjointMethod*($F, M, \epsilon^{(0)}, \delta, N_{max}$)**:**

$$\underset{\epsilon}{\text{maximize}} \quad F[\psi(\mathbf{x}), \epsilon(\mathbf{x})] \quad \text{subject to} \quad M[\psi(\mathbf{x}), \epsilon(\mathbf{x})] = 0$$

$\epsilon^{(k)} \leftarrow \epsilon^{(0)}$; $k \leftarrow 0$; $F_{\text{old}} \leftarrow 0$;
repeat
  $\psi^{(k)} \leftarrow \texttt{ForwardSolve}(\epsilon^{(k)})$;
  $\frac{\delta F}{\delta \psi} \leftarrow \texttt{AdjointSourceTerm}(F, \psi^{(k)})$;
  $\lambda^{(k)} \leftarrow \texttt{AdjointSolve}(\frac{\delta F}{\delta \psi}, M, \psi^{(k)})$;
  $\frac{\delta F}{\delta \epsilon} \leftarrow \frac{\partial F}{\partial \epsilon} - \lambda^{(k)} \cdot \frac{\delta M}{\delta \epsilon}$;
  $\epsilon^{(k+1)} \leftarrow \texttt{UpdateDesign}(\frac{\delta F}{\delta \epsilon}, \epsilon^{(k)})$;
  $F_{\text{new}} \leftarrow \texttt{EvaluateObjective}(F, \psi^{(k)}, \epsilon^{(k)})$;
  $k \leftarrow k + 1$;
until $|F_{new} - F_{old}| < \delta$ ***or*** $k \geq N_{max}$
$\epsilon^* \leftarrow \epsilon^{(k)}$;
return ϵ^* ;

In NEOCNN, the digitized meta-structure of nanoswitch has a region of $2.4 \times 6\ \mu m^2$, with the grid scale at 32×80. Each grid contains a possible nanohole with a diameter of $60nm$ that could be etched during fabrication according to the topology optimization outcome.

The s-parameter matrix can be used to describe the behavior of an optical component, which summarizes the inner relationship of input and output as $O_{1\times2} = I_{1\times2} \cdot \mathbf{s}_{2\times2}$. In the context of an optical butterfly, a nanoswitch has a s-matrix of

$$\mathbf{s} = \frac{1}{\sqrt{1+2^{2n}}} \begin{bmatrix} 1 & 1 \\ 2^n & -1 \end{bmatrix}, (n = 1, 2, 3, 4) \tag{3}$$

, which could be implemented with a single nanoswitch. One of the possible nanohole distributions in NEOCNN is demonstrated in Fig. 3 when $n = 1$.

Unlike the design in [52], each optical processor operates differently and has to be designed individually, the nanoswitches in the optical NTT mesh are configured to modulate with a limited number of different n, which greatly reduces the design overhead. Moreover, the generic mesh design in NEOCNN enables the inference of various network models, rather than having the weight fixed after fabrication.

Additionally, the application of the Fermat number as the modulo q enables the simplification of the transfer matrix in the NTT process, which could make sure that the error accumulation is minimized and chances for error could be eliminated.

3.5 Optical NTT Mesh

NTT could be implemented with cascaded optical butterflies, the arithmetic function of which is depicted in Fig. 4a. Compared to electrical NTT modules that recursively utilize the butterfly, the optical butterfly mesh operates in a feed-forward manner, allowing on-the-fly NTT and avoiding the repetitive shuttling of data inherent to recursive approaches.

Each node within the mesh is constructed using a pre-coded nanoswitch, which directs the incoming light to the appropriate output ports. These nanoswitches are the core components that enable the reconfigurability of the optical paths and, consequently, the arithmetic operations of the NTT. The feed-forward nature of the mesh enables the NTT transform to be conducted at the speed of light and a frequency of 10 GHz or even higher. The passivity of photonic components is another significant advantage, as they require no energy to maintain the state of light passing through them. This characteristic of the mesh results in a process free from energy consumption during NTT operations. Moreover, the scale of the mesh is designed to obviate the need for additional energy compensation.

An 8-point radix-2 NTT with a Fermat number as modulo $q = F_n$ is illustrated in Figure 5 as an example. This straightforward implementation of the Fast NTT algorithm consists of $log(n)$ stages, each containing $n/2$ butterfly modules.

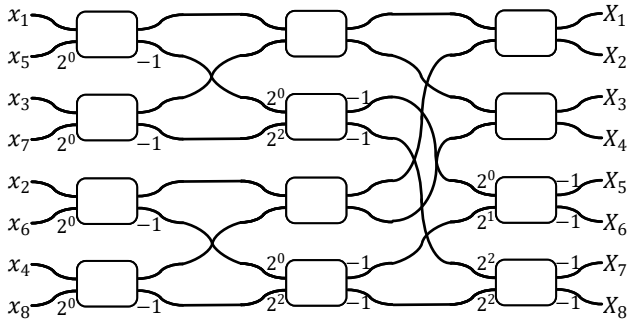


Figure 5: Nanoswitch mesh for NTT/INTT

3.6 Hardware Architecture and Dataflow

As depicted in Figure 6, the NEOCNN comprises several parts, including the signal generation module, the NTT mesh, the reduction module, the Hadamard product module, and the memory attached. The system's modular nature allows for flexibility in handling various convolution sizes and enables scalability for larger networks or datasets.

The input signal is created by a laser which emits light that encodes information. This information-laden light is then directed through the NTT mesh, which transforms the signal into the NTT

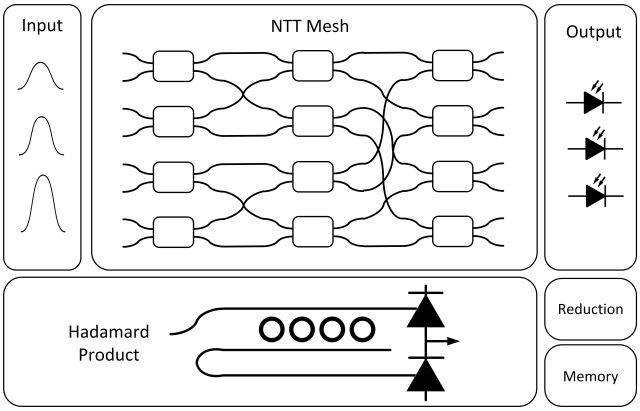


Figure 6: System Overview

domain for efficient convolution operations. The NTT mesh functions passively, meaning it relies on pre-coded configurations that correspond to mathematical operations necessary for the transformation, thus eliminating the need for constant data movement and reducing the computational load on the system's memory.

The Hadamard product module then performs an element-wise multiplication. This is done using a silicon weight bank that imprints the network's learned parameters onto the optical signal, akin to the design proposed in MR-based ONNs.

3.7 Functional Modules

The NEOCNN's computational efficiency is greatly enhanced through the use of several specialized modules. Notably, the Hadamard product, the dot-product operation unit in the network, is facilitated by silicon weight banks—technology traditionally associated with MR-based systems. Precision in this operation is paramount and is achieved by adjusting the MR modulator's oscillation wavelength and fine-tuning input signals, which was previously discussed in Section 2.1.

With the multiple data reuse patterns in CNN, the free spectral range (FSR)-parallelism could be applied in this Hadamard product unit as well. As depicted in the transmission spectrum in Fig. 2b, the tuning of MR can operate on multiple wavelengths simultaneously with their interval at an integer multiple of the FSR. This approach cleverly utilizes wavelength-division-multiplexing (WDM) to enhance system parallelism—potentially increasing throughput by a factor of four or more.

The reduction unit of the NEOCNN, which is integral for simplifying the information post-transformation, is implemented using electronic circuits. This choice is strategic; by employing Fermat numbers as the moduli, the complexity of the reduction process is significantly diminished. Unlike the general Number Theoretic Transform (NTT) scheme, this method necessitates only simple bit operations. As a result, the footprint and energy consumption of the reduction unit is minimal—rendering it virtually cost-free and occupying an inconsequential portion of the system's resources. This design choice underscores the NEOCNN's commitment to efficiency and scalability.

Table 2: Photonic components parameters

	Component	Parameter	Spec†	Power (W)	Area (mm^2)
O*	Nanoswitch Mesh	Number	32	/	0.12
		Size	8×4		
		Ins Loss	-0.8dB		
		Crosstalk	-20dB		
	MR weight bank[13]	Number	32	< 0.01	0.20
		Size	16		
		Precision	4-bit		
	Laser[43]	Number	16×4	0.64	3.00
	PD[30]	Frequency	10GHz	1.28	1.28
		Number	512		
E*	Reduction Unit	Number	64	< 0.01	0.01
	Shifter Adder	Number	64	0.54	0.16
OE*	ADC[26]	Resolution	8 bits	3.79	1.46
		Number	16 × 32		
	DAC[14]	Resolution	4 bits	0.12	0.07
		Number	32×32		
	Total			**6.37**	**6.68**

† 16-point NTT, 32 meshes, FSR_level = 4
∗ O: Optical components; E: Electronic components; OE: OE interface

4 SIMULATION AND EXPERIMENTAL VERIFICATION

In this section, a detailed performance analysis is conducted through power, latency, and throughput. Comparison is performed against SOTA OCNN, FPGA-based CNN accelerators, and ASIC. Besides, the crosstalk and loss, which may significantly impact the accuracy of CNN inference, have also been thoroughly analyzed.

4.1 Evaluation Setup

As depicted in Table 2, the NEOCNN architecture is composed of 32 nanoswtich meshes, each mesh intricately designed to execute a 16-point NTT. MR weight banks are responsible for the Hadamard product in the NTT domain.

The functionality of optical components is evaluated through the photonic device modeling and simulation environment BOSIM[9], which provides a comprehensive platform to simulate the behavior of photonic circuits and guarantees that the NEOCNN's components adhere to their expected performance specifications. The systematic synthesis analysis is performed using a heterogeneous system simulation platform JADE [22], which enables the evaluation of the NEOCNN in a structured manner, taking into account various system-level considerations such as power, latency, and throughput.

Table 3: Comparison with other NTT accelerators

Design	Platform	Method	Number of PEs	Bitwidth	Frequency (MHz)	Model	Latency (ms)	Throughput (GOPS)	Power Efficiency (GOPS/W)
Ours	Photonic	NTT	32	8	10k	VGG-16 GoogLeNet	**1.95** **0.70**	**61020** **15096**	**9579.3** **2369.9**
[46]	Photonic	Spatial	1	8	62.9k	/	/	11300	NA
[18]	FPGA Alveo U50	NTT	2048	8 / 21	200	VGG-16 GoogLeNet	13.9 3.58	2859.5 990.3	110.0 38.1
[45]	FPGA VX485t	NTT	/	16	150	VGG-16	150.2	264.6	/
[27]	FPGA ZCU 102	Winograd	2345	8 ~ 16	214	VGG-16 Yolo-v2	19.67 13.9	3120.3 805.6	NA
[20]	FPGA XC7VX980T	Spatial	3395	8 / 16	150 100	VGG-16 ResNet-101	NA	1000 600	69.64 48.43
[21]	FPGA XCVC1902	Spatial	128	8	1333	VGG-16	16.99	10952	303.4
[25]	ASIC	Spatial	16	16	1000	NA	NA	1056	1771.8
[39]	RRAM	Spatial	8	8	1000	NA	NA	1707	25.9

4.2 Overall Performance Comparison

The performance of NEOCNN is evaluated by comparing it with other state-of-the-art (SOTA) CNN accelerators. Table 3 provides a comprehensive comparison in terms of key metrics such as processing elements (PEs), bit-width, operating frequency, models supported, latency, throughput, and power efficiency.

NEOCNN surpasses SOTA electrical accelerators in terms of throughput, power efficiency, and latency comprehensively, achieving an unparalleled throughput of 61$TOPS$ and power efficiency of 9.6 $TOPS/W$. Thanks to the ultra-high speed of optical devices, NEOCNN could minimize the inference efficiency. Notably, the area is 62% smaller with the adoption of the nanoswitch, which is not explicitly indicated in the table.

The NTT-based convolution method also contributes to the flexibility of NEOCNN, which enables the inference of popular models such as VGG-16 and GoogLeNet. Though the flexibility is still limited compared to that of the electrical accelerators (throughput degradation in GoogLeNet), this flexibility of the network provides great convenience in system scalability compared to other ONNs [46, 52] that only fit with dedicated models.

In conclusion, NEOCNN demonstrates exceptional performance characteristics in comparison to other state-of-the-art accelerators. The cutting-edge photonic technology, combined with the notable operating frequency, enables both high throughput and low latency, which are critical for real-time CNN applications. Moreover, the remarkable power efficiency as well as the flexibility of the NEOCNN is indicative of its potential for deployment in extremely computation-intensive scenarios.

4.3 Energy and Area Breakdown

The energy and area breakdown for NEOCNN provides insights into the efficiency and spatial distribution of its components, as depicted in Figure 7.

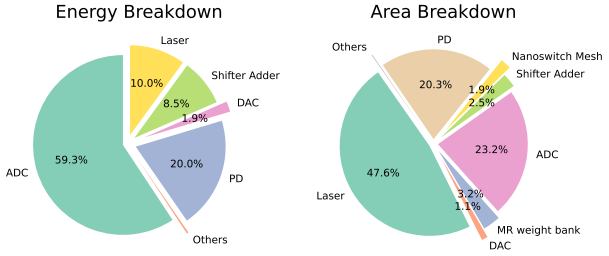


Figure 7: Energy and Area Breakdown for NEOCNN

NEOCNN's design primarily focuses on minimizing the energy consumption associated with optical-electrical (O-E) and electrical-optical (E-O) conversions, which are traditionally the main contributors to power usage in optical convolutional neural network (OCNN) systems. By taking advantage of the passivity and the multiple parallelism, NEOCNN achieves a considerable improvement in energy efficiency over previous OCNN architectures.

In regards to the area, the NEOCNN's layout is optimized through the integration of a nanoswitch mesh, which constitutes the largest portion of the area. This mesh replaces the bulkier Mach-Zehnder Interferometer (MZI) arrays found in traditional designs, leading to a more compact and area-efficient footprint without compromising performance.

As for the energy breakdown, photodetectors (PD) and analog-to-digital converters (ADC) are the major consumers, each accounting for more than 20% of the total energy. The energy usage of lasers and digital-to-analog converters (DAC) is remarkably low, indicating optimized conversion efficiency. The shifter adder and other miscellaneous components represent minor energy expenditures within the system.

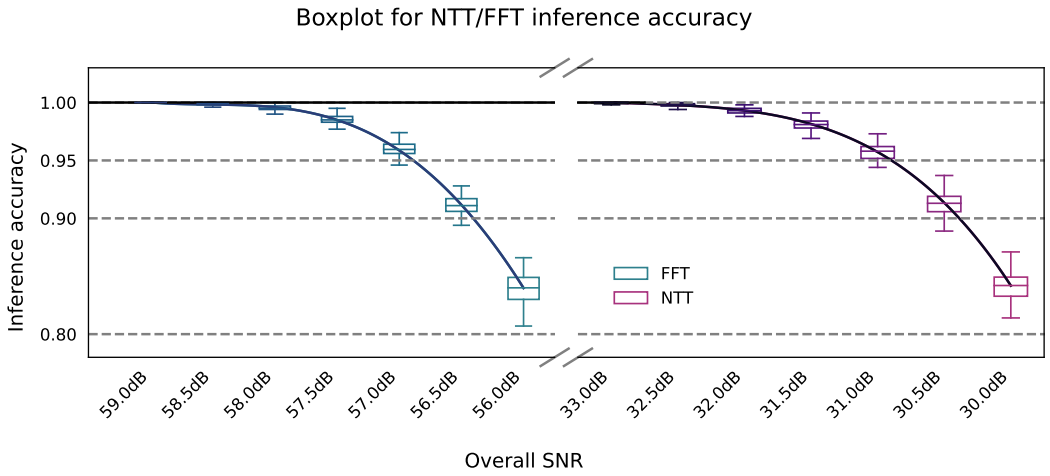


Figure 8: Monte-Carlo analysis of SNR's influence on the NTT/FFT inference

4.4 Analysis of Noise, Crosstalk, and Error

Compared to the traditional spatial convolution computing paradigms, either NTT-based or FFT-based convolution methods necessitate a heightened computational precision due to the analog nature of the optical accelerator. This precision is critical, as any error during the inference of the transform can lead to substantially incorrect convolution results. Thus, a comprehensive analysis of noise, crosstalk, and error is completed on the accuracy of the NEOCNN.

Signal-to-Noise Ratio (SNR): As the main computing component of the whole system, the nanoswitch is configured to a 4-bit modulation by default. This modest bit precision is intentional, allowing for a higher tolerance to modulation imperfections and noise. As depicted in Figure 8, the SNR for NTT-based convolution accuracy begins to degrade at approximately 32dB. This performance is notably superior to that of the FFT-based method, which experiences degradation at 58dB under the same hardware conditions.

The primary source of inference noise in ONNs is the thermal noise, a byproduct of intense on-chip electro/thermal modulation. NEOCNN, in contrast to conventional OCNN systems, requires real-time modulation only in a small fraction of its core computation unit, specifically in the Hadamard product. This design choice contributes to a lower noise inference environment for NEOCNN, which boasts a reasonable SNR requirement even when compared to conventional OCNNs with SNRs as low as 30dB [29].

Crosstalk: Crosstalk effects have been thoroughly investigated. Because of the integer arithmetic nature of NTT operations, the theoretic crosstalk upper bound could be deducted given certain parameter settings. In NEOCNN's default setting, the crosstalk upper bound of NTT-based convolution is approximately $-27.1dB$. As a comparison, FFT-based convolution exhibits a theoretical crosstalk level of around −66.4dB.

Advanced MZI modulators can achieve crosstalk as low as −60dB[44], yet the system-wide crosstalk for ONNs should be limited to between −20dB and −30dB. This range accounts for the cumulative crosstalk introduced by each modulation event, which necessitates a more stringent crosstalk requirement. NEOCNN sidesteps many of these concerns with its pre-coded NTT inference and absence of EO conversion, leading to simpler nanoswitch fabrication. Through the application of isolation and optimization techniques, NEOCNN operates effectively below the crosstalk threshold.

Loss: Optical loss is another critical factor affecting OCNN accuracy and is rigorously analyzed as shown in Figure 9. Employing a 4-layer CNN (comprising 2 convolutional layers and 2 MLPs) tested on the MNIST dataset, the system's accuracy was evaluated under various crosstalk and loss conditions. NEOCNN demonstrates a high degree of error tolerance due to its relaxed parameter settings. Even with an optical loss as high as −0.4dB — a highly improbable scenario—the system's inference accuracy remains impressively stable at 92.4%.

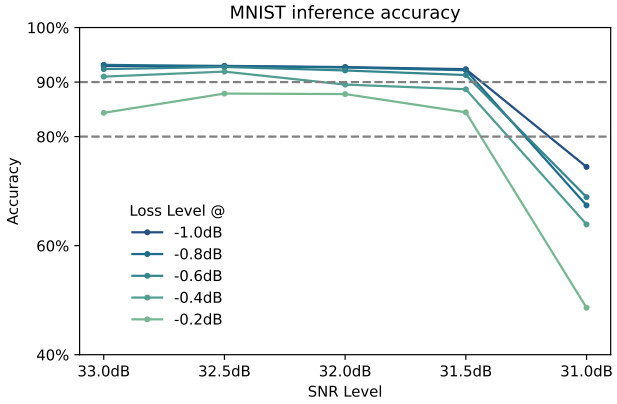


Figure 9: MNIST inference accuracy under various loss and SNR settings.

We also provide a thorough investigation into the influence of all potential noise factors on the inference accuracy of various network models. The models included are LeNet-5, AlexNet, VGG-16, DenseNet, and MobileNet. Each model was subjected to a consistent experiment setup, which involved training under default loss conditions with an added crosstalk and noise setting, to simulate realistic operational environments.

The models were trained with a batch size of 640 and for a total of 10 epochs. The accuracy results, presented in Table 4, demonstrate that the performance across different architectures is relatively stable, thereby underscoring the robustness of NEOCNN.

Table 4: Accuracy of models

	LeNet-5	AlexNet	VGG-16	DenseNet	MobileNet
NEOCNN	93.85%	92.75%	93.87%	94.26%	91.13%
GPU	94.25%	95.72%	95.37%	95.64%	92.26%

5 CONCLUSION

In conclusion, the NEOCNN framework pioneers the integration of NTT into optical convolutional neural network accelerators, heralding a significant leap in computational efficiency for ONNs. The novel photonic butterfly structure, in conjunction with an NTT mesh, realizes on-the-fly NTT transformations that drastically reduce both computational resource consumption and physical hardware footprint. Our rigorous analysis demonstrates the NEOCNN's resilience against potential crosstalk and loss factors, with the system showing outstanding inference accuracy.

The NEOCNN achieves a groundbreaking throughput of 61 Tera Operations per Second (TOPs) while maintaining an impressive power efficiency of 9.6 TOPs/Watt during VGG-16 inference tasks. These figures not only surpass state-of-the-art electronic accelerators but also advance the capabilities of existing ONN models. This work sets a new benchmark for future research in optical computing and showcases the transformative potential of merging number theoretical transforms with optical computing paradigms for neural network acceleration.

ACKNOWLEDGMENTS

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0013).

REFERENCES

[1] [n. d.]. DALL·E: Creating Images from Text.
[2] [n. d.]. Introducing ChatGPT.
[3] Salma Afifi, Febin Sunny, Mahdi Nikdast, and Sudeep Pasricha. 2023. TRON: Transformer Neural Network Acceleration with Non-Coherent Silicon Photonics. In *GLSVLSI*. 15–21.
[4] RC Agarwal and C Burrus. 1974. Fast Convolution Using Fermat Number Transforms with Applications to Digital Filtering. *IEEE Transactions on Signal Processing* 22, 2 (1974), 87–97.
[5] Rashmi Agrawal, Leo de Castro, Guowei Yang, Chiraag Juvekar, Rabia Yazicigil, Anantha Chandrakasan, Vinod Vaikuntanathan, and Ajay Joshi. 2023. FAB: An FPGA-based Accelerator for Bootstrappable Fully Homomorphic Encryption. In *HPCA*. IEEE, 882–895.
[6] Amir H Atabaki, Sajjad Moazeni, Fabio Pavanello, Hayk Gevorgyan, Jelena Notaros, Luca Alloatti, Mark T Wade, Chen Sun, Seth A Kruger, Huaiyu Meng, et al. 2018. Integrating Photonics with Silicon Nanoelectronics for the next Generation of Systems on a Chip. *Nature* 556, 7701 (2018), 349–354.
[7] Hengameh Bagherian, Scott Skirlo, Yichen Shen, Huaiyu Meng, Vladimir Ceperic, and Marin Soljacic. 2018. On-Chip Optical Convolutional Neural Networks. *arXiv preprint arXiv:1808.03303* (2018). arXiv:1808.03303
[8] Utsav Banerjee, Abhishek Pathak, and Anantha P Chandrakasan. [n. d.]. 2.3 An Energy-Efficient Configurable Lattice Cryptography Processor for the Quantum-Secure Internet of Things. In *ISSCC*.
[9] Xuanqi Chen, Zhifei Wang, Yi-Shing Chang, Jiang Xu, Jun Feng, Peng Yang, Zhehui Wang, and Luan H. K. Duong. 2020. Modeling and Analysis of Optical Modulators Based on Free-Carrier Plasma Dispersion Effect. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39, 5 (May 2020), 977–990.
[10] Yitong Chen, Maimaiti Nazhamaiti, Han Xu, Yao Meng, Tiankuang Zhou, Guangpu Li, Jingtao Fan, Qi Wei, Jiamin Wu, Fei Qiao, et al. 2023. All-Analog Photoelectronic Chip for High-Speed Vision Tasks. *Nature* (2023), 1–10.
[11] Yu-Hsin Chen, Tushar Krishna, Joel S Emer, and Vivienne Sze. 2016. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE journal of solid-state circuits* 52, 1 (2016), 127–138.
[12] Robert H Dennard, Fritz H Gaensslen, Hwa-Nien Yu, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. 1974. Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions. *IEEE Journal of solid-state circuits* 9, 5 (1974), 256–268.
[13] Raphaël Dubé-Demers, Sophie LaRochelle, and Wei Shi. 2016. Ultrafast Pulse-Amplitude Modulation with a Femtojoule Silicon Photonic Modulator. *Optica* 3, 6 (June 2016), 622–627.
[14] Hossein Eslahi, Tara J. Hamilton, and Sourabh Khandelwal. 2022. Ultra Compact and Linear 4-Bit Digital-to-Analog Converter in 22nm FDSOI Technology. In *ISCAS*. 2778–2781.
[15] Michael Y-S Fang, Sasikanth Manipatruni, Casimir Wierzynski, Amir Khosrowshahi, and Michael R DeWeese. 2019. Design of Optical Neural Networks with Component Imprecisions. *Optics Express* 27, 10 (2019), 14009–14029.
[16] Johannes Feldmann, Nathan Youngblood, C David Wright, Harish Bhaskaran, and Wolfram HP Pernice. 2019. All-Optical Spiking Neurosynaptic Networks with Self-Learning Capabilities. *Nature* 569, 7755 (2019), 208–214.
[17] Jiaqi Gu, Zheng Zhao, Chenghao Feng, Zhoufeng Ying, Mingjie Liu, Ray T Chen, and David Z Pan. 2020. Toward Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability. *IEEE TCAD* 40, 9 (2020), 1796–1809.
[18] Seongmin Hong, Yashael Faith Arthanto, Joo-Young Kim, et al. 2022. Accelerating Deep Convolutional Neural Networks Using Number Theoretic Transform. *IEEE TCAS-I* 70, 1 (2022), 315–326.
[19] Jie Huang, Hansi Ma, Dingbo Chen, Huan Yuan, Jinping Zhang, Zikang Li, Jingmin Han, Jiagui Wu, and Junbo Yang. [n. d.]. Digital Nanophotonics: The Highway to the Integration of Subwavelength-Scale Photonics: Ultra-compact, Multi-Function Nanophotonic Design Based on Computational Inverse Design. *Nanophotonics* 10, 3 ([n. d.]), 1011–1030.
[20] Wenjin Huang, Huangtao Wu, Qingkun Chen, Conghui Luo, Shihao Zeng, Tianrui Li, and Yihua Huang. 2021. FPGA-based High-Throughput CNN Hardware Accelerator with High Computing Resource Utilization Ratio. *IEEE Transactions on Neural Networks and Learning Systems* 33, 8 (2021), 4069–4083.
[21] Xijie Jia, Yu Zhang, Guangdong Liu, Xinlin Yang, Tianyu Zhang, Jia Zheng, Dongdong Xu, Zhuohuan Liu, Mengke Liu, Xiaoyang Yan, et al. 2022. XVDPU: A High Performance CNN Accelerator on Versal Platform Powered by AI Engine. *ACM Transactions on Reconfigurable Technology and Systems* (2022).
[22] Fan Jiang, Rafael KV Maeda, Jun Feng, Shixi Chen, Lin Chen, Xiao Li, and Jiang Xu. 2022. Fast and Accurate Statistical Simulation of Shared-Memory Applications on Multicore Systems. *IEEE Trans Parallel Distrib Syst* 33, 10 (2022), 2455–2469.
[23] Dai Li, Akhil Pakala, and Kaiyuan Yang. 2022. MeNTT: A Compact and Efficient Processing-in-Memory Number Theoretic Transform (NTT) Accelerator. *VLSI* 30, 5 (2022), 579–588.
[24] Xianbin LI, Jiaqi LIU, Yuying ZHANG, and et. al. [n. d.]. PhotonNTT: Energy-efficient Parallel Photonic Number Theoretic Transform Accelerator. In *DATE2024*.
[25] Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Teman, Xiaobing Feng, Xuehai Zhou, and Yunji Chen. 2015. Pudiannao: A Polyvalent Machine Learning Accelerator. *ACM SIGARCH Computer Architecture News* 43, 1 (2015), 369–381.
[26] Juzheng Liu, Mohsen Hassanpourghadi, and Mike Shuo-Wei Chen. 2022. A 10GS/s 8b 25fJ/c-s 2850um2 Two-Step Time-Domain ADC Using Delay-Tracking Pipelined-SAR TDC with 500fs Time Step in 14nm CMOS Technology. In *ISSCC*, Vol. 65. 160–162.
[27] Xinheng Liu, Yao Chen, Cong Hao, Ashutosh Dhar, and Deming Chen. [n. d.]. WinoCNN: Kernel Sharing Winograd Systolic Array for Efficient Convolutional Neural Network Acceleration on FPGAs. In *ASAP*.

[28] Yinyi Liu, Jiaxu Zhang, Jun Feng, Shixi Chen, and Jiang Xu. 2022. Reduce Footprints of Multiport Interferometers by Cosine-Sine-Decomposition Unfolding. In *OFC*. W2A–4.

[29] Yinyi Liu, Jiaxu Zhang, Jun Feng, Shixi Chen, and Jiang Xu. 2022. A Reliability Concern on Photonic Neural Networks. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 1059–1064.

[30] Ping Ma, Yannick Salamin, Benedikt Baeuerle, Arne Josten, Wolfgang Heni, Alexandros Emboras, and Juerg Leuthold. 2019. Plasmonically Enhanced Graphene Photodetector Featuring 100 Gbit/s Data Reception, High Responsivity, and Compact Size. *ACS Photonics* 6, 1 (Jan. 2019), 154–161.

[31] Simei Mao, Lirong Cheng, Fasial Nadeem Khan, Zihan Geng, Qian Li, and HY Fu. [n. d.]. Inverse Design of High-Dimensional Nanostructured 2× 2 Optical Processors Based on Deep Convolutional Neural Networks. *Journal of Lightwave Technology* ([n. d.]).

[32] Gordon Moore. 2021. Cramming More Components onto Integrated Circuits (1965). (2021).

[33] Hamid Nejatollahi, Saransh Gupta, Mohsen Imani, Tajana Simunic Rosing, Rosario Cammarota, and Nikil Dutt. 2020. Cryptopim: In-memory Acceleration for Lattice-Based Cryptographic Hardware. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[34] Henri J Nussbaumer and Henri J Nussbaumer. 1982. *The Fast Fourier Transform*. Springer.

[35] Yongmo Park, Ziyu Wang, Sangmin Yoo, and Wei D Lu. 2022. RM-NTT: An RRAM-Based Compute-in-Memory Number Theoretic Transform Accelerator. *IEEE J. Explor* 8, 2 (2022), 93–101.

[36] Nicola Peserico, Russell Schwartz, Hangbo Yang, Xiaoxuan Ma, Mostafa Hosseini, Puneet Gupta, Hamed Dalir, and Volker J Sorger. 2022. FFT-based Convolution Neural Network on Silicon Photonics Platform. In *2022 IEEE Photonics Conference (IPC)*. IEEE, 1–2.

[37] Nikola Samardzic, Axel Feldmann, Aleksandar Krastev, Srinivas Devadas, Ronald Dreslinski, Christopher Peikert, and Daniel Sanchez. 2021. F1: A Fast and Programmable Accelerator for Fully Homomorphic Encryption. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*. 238–252.

[38] SEAL 2023. Microsoft SEAL (Release 4.1).

[39] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. [n. d.]. ISAAC: A Convolutional Neural Network Accelerator with in-Situ Analog Arithmetic in Crossbars. *ACM SIGARCH Computer Architecture News* ([n. d.]).

[40] Yichen Shen, Nicholas C Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, et al. 2017. Deep Learning with Coherent Nanophotonic Circuits. *Nature Photonics* 11, 7 (2017), 441.

[41] Shiming Song, Wei Tang, Thomas Chen, and Zhengya Zhang. 2018. LEIA: A 2.05 Mm 2 140mW Lattice Encryption Instruction Accelerator in 40nm CMOS. In *CICC*. IEEE, 1–4.

[42] Chen Sun, Mark T Wade, Yunsup Lee, Jason S Orcutt, Luca Alloatti, Michael S Georgas, Andrew S Waterman, Jeffrey M Shainline, Rimas R Avizienis, Sen Lin, et al. 2015. Single-Chip Microprocessor That Communicates Directly Using Light. *Nature* 528, 7583 (2015), 534–538.

[43] Zhenxing Sun, Zhirui Su, Rulei Xiao, Yaguang Wang, Kui Liu, Feng Wang, Yashe Liu, Tao Fang, Yi-Jen Chiu, and Xiangfei Chen. 2022. Tunable Laser via High-Density Integration of DFB Lasers with High Precision Wavelength Spacings. *IEEE Photonics Technology Letters* 34, 9 (2022), 467–470.

[44] Callum M Wilkes, Xiaogang Qiang, Jianwei Wang, Raffaele Santagati, Stefano Paesani, Xiaoqi Zhou, David AB Miller, Graham D Marshall, Mark G Thompson, and Jeremy L O'Brien. 2016. 60 dB High-Extinction Auto-Configured Mach–Zehnder Interferometer. *Optics letters* 41, 22 (2016), 5318–5321.

[45] Weihong Xu, Xiaohu You, and Chuan Zhang. 2017. Using Fermat Number Transform to Accelerate Convolutional Neural Network. In *ASICON*. IEEE, 1033–1036.

[46] Xingyuan Xu, Mengxi Tan, Bill Corcoran, Jiayang Wu, Andreas Boes, Thach G Nguyen, Sai T Chu, Brent E Little, Damien G Hicks, Roberto Morandotti, et al. 2021. 11 TOPS Photonic Convolutional Accelerator for Optical Neural Networks. *Nature* 589, 7840 (2021), 44–51.

[47] Nathan Youngblood. 2022. Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication. *IEEE journal of selected topics in quantum electronics : a publication of the IEEE Lasers and Electro-optics Society* 29 (2022), 1–11.

[48] Huan Yuan, Zhicheng Wang, Zheng Peng, Jiagui Wu, and Junbo Yang. 2023. Ultra-Compact and NonVolatile Nanophotonic Neural Networks. *Advanced Optical Materials* 11, 16 (2023), 2300215.

[49] Cong Zhang, Dongsheng Liu, Xingjie Liu, Xuecheng Zou, Guangda Niu, Bo Liu, and Quming Jiang. 2021. Towards Efficient Hardware Implementation of NTT for Kyber on FPGAs. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

[50] Weipeng Zhang, Chaoran Huang, Hsuan-Tung Peng, Simon Bilodeau, Aashu Jha, Eric Blow, Thomas Ferreira De Lima, Bhavin J Shastri, and Paul Prucnal. 2022. Silicon Microring Synapses Enable Photonic Deep Learning beyond 9-Bit Precision. *Optica* 9, 5 (2022), 579–584.

[51] Yuying Zhang, Sarveswara Reddy Sathi, Zili Kou, Sharad Sinha, and Wei Zhang. 2023. Tensor-Product-Based Accelerator for Area-Efficient and Scalable Number Theoretic Transform. In *FCCM*. IEEE, 174–183.

[52] Caiyue Zhao, Jiguang Wang, Simei Mao, Xuanyi Liu, Wai Kin, Victor Chan, and HY Fu. 2023. End-to-End Optimization for a Compact Optical Neural Network Based on Nanostructured 2× 2 Optical Processors. *IEEE Photonics Journal* (2023).

[53] Hailong Zhou, Jianji Dong, Junwei Cheng, Wenchan Dong, Chaoran Huang, Yichen Shen, Qiming Zhang, Min Gu, Chao Qian, Hongsheng Chen, et al. 2022. Photonic Matrix Multiplication Lights up Photonic Accelerator and Beyond. *Light: Science & Applications* 11, 1 (2022), 30.