A Reliability Concern on Photonic Neural Networks

Yinyi Liu[1] , Jiaxu Zhang[1], Jun Feng[1], Shixi Chen[1], Jiang Xu[2,1,†]
[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
[2]Microelectronics Thrust, The Hong Kong University of Science and Technology
[†]*Corresponding author: jiang.xu@ust.hk*

***Abstract*—Emerging integrated photonic neural networks have experimentally proved to achieve an ultra-high speedup of deep neural network training and inference in the optical domain. However, photonic devices suffer from the inherent crosstalk noise and loss, inevitably leading to reliability concerns. This paper systematically analyzes the impacts of crosstalk and loss on photonic computing systems. We propose a crosstalk-aware model for reliability estimation and find out the worst-case bounds as we increase the footprints and scales of the photonic chips. Our evaluations show that -30dB crosstalk noise can cause maximal photonic chip integration to a sharp drop by 109x. To facilitate very-large-scale photonic integration for future computing, we further propose multiple heterogeneous bijou photonic-cores to address the crosstalk-aware reliability concern.**

***Index Terms*—reliability, crosstalk, photonic neural networks, photonic cores, many-core system, heterogeneous chiplet, VLSI**

I. Introduction

Integrated photonic neural networks (PNNs) have been demonstrated remarkably adept at accelerating deep neural network (DNN) training and inference processes, with ultra-high speed up to 10^{12} multiply-accumulate (MAC) operations per second [1] and ultra-low power consumption of 10^{-9} Watt per switching component [2], [3]. As shown in Fig. 1, a state-of-the-art PNN core contains a phase controller, several photodetectors and programmable unitary blocks. Matrix multiplications in DNN layers can be reinterpreted to unitary operations by applying the singular value decomposition (SVD) method, and finally deployed on PNN chips by controlling phase shifting [4], [5]. Operating in analogue paradigm, PNNs take great advantages of the photonic nature to proceed signals, and avoid the limitations imposed by time and power consumed due to frequent memory access for data storage.

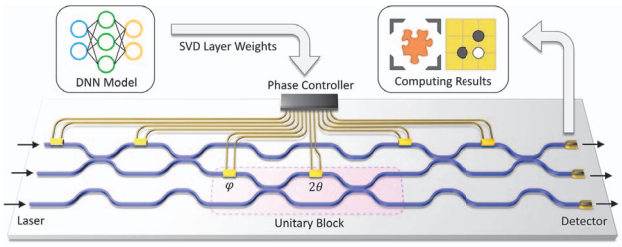


Fig. 1. Schematic of the Photonic Neural Networks. MAC operations in DNNs are converted into unitary operations, run on the unitary blocks, performed at the speed of light and detected at the rate over 200GHz [6].

In contrast to the way of digital processing in conventional electronics, the essence of PNN is to make full use of analogue calculation. This paradigm of the aforementioned photonic computing system has proved its *effectiveness*. It boosts the computational density tremendously. However, this comes at a cost. In general, analogue fails in signal integrity contrarily to digital when it comes to very-large-scale (VLS) systems. To what extent users can trust the results produced by the PNN system becomes an issue that needs urgent investigation, leading to *reliability* concerns. Empirically, four predominant factors affect the system reliability of PNN:

1) **Loss**: After injecting laser into PNNs, the system can be passive, and suffers absorption and radiation during signal propagating along the waveguide, which incurs loss.
2) **Crosstalk**: Disturbances such as thermal fluctuation or coupling drift cause a small portion of the power to be directed to the unexpected port and become noises, when the light passes through an intersect or coupling region.
3) **Finite Precision**: To modulate and demodulate the data onto and from PNNs, digital-to-analog (DA) and analog-to-digital (AD) are essential. The quantization subject to finite precision depth incurs slight deviation.
4) **Amplifier Linearity**: To well compensate the attenuation caused by loss, amplifiers are applied especially in the VLS-PNNs. Their defects of linearity and dynamic ranges further incur distortion of the photonic computing results.

Prior works have systematically discussed the problems about finite precision [7], [8], and further developed error-tolerant mixed precision techniques [9] and adaptive quantization [10]. Researchers in the device domain make efforts to extend the dynamic ranges and to improve the linearity of amplifiers [11]. The influence of loss and the optimized design has also been reported [12]. But for the crosstalk, there are few systematic analyses of photonic computing systems.

Crosstalk is never a trivial issue. The loss-only model is completely inadequate. Previous efforts on manufacturing PNN prototypes focused on small-scale circuits, and at this scale, the impacts of crosstalk had not yet been conspicuously revealed. As we are towards VLS-PNNs, the accumulated crosstalk will explode with the increase of the circuit scale. That is to say, when the circuit is realized at a relative small scale, loss is the primary factor. But when growing the circuit scale, crosstalk level becomes the key factor to determine the PNN reliability.

Here, we present a theoretical model to reveal the impacts of crosstalk and loss on PNN systems. Experiments consisting of representative deep learning applications is shown to corroborate the severity of reliability degradation due to crosstalk. To address the crosstalk-aware reliability concern, we look deeper into the PNNs and present the scheme of the multiple heterogeneous bijou photonic cores (PHOEBE) architecture.

The remainder of this paper is organized as follows. Section II formulates the crosstalk model of PNNs. Section III benchmarks the test-accuracy degradation of DNNs caused by crosstalk. Section IV describes the Phoebe architecture and discusses its superiority. Finally, Section V concludes this work.

II. Formulate Crosstalk in PNN Systems

In this section, we elucidate the mechanism of crosstalk from the physical point of view with validation in *Lumerical*, and extract optical parameters. Then we formulate the crosstalk model for photonic neural networks and analyze the bounds.

A. Mechanism of Crosstalk Noise

Crosstalk is an intrinsic characteristic of photonic devices within waveguide networks [13], as shown in Fig. 2. During propagating in waveguides, light is likely to encounter the coupling drift due to modulation bias, or phase vagueness due to scattering on rough sidewalls [14]. Consequently, part of the signal power leaks into unexpected paths or ports and finally becomes noise. Furthermore, crosstalk can be distinguished into coherent crosstalk and incoherent crosstalk.

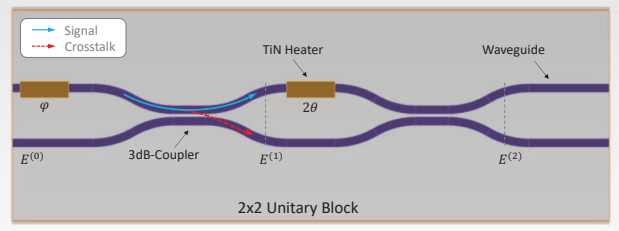


Fig. 2. A 2×2 Unitary Block comprises two Mach-Zehnder Interferometers (MZI) and two metal heaters for thermo-optic tuning, layout from *Lumerical*. Fabrication imperfectness of the coupling regions incurs the crosstalk noise.

Modulation bias deviates the phase of light from the correct position while maintaining light coherence. It incurs coherent crosstalk, which attenuates the signal power. The proof is shown as follows: take the original signal as the reference phase, the coherent crosstalk induces a phase shift β,

$$E_s \rightarrow (1-K)E_s + KE_s e^{i\beta} \tag{1}$$

where K is the coherent crosstalk coefficient, E_s indicates the optical field of signals. $|KE_s|$ becomes the coherent crosstalk.

$$\begin{aligned} |E_s'| &= \sqrt{(1-K)^2|E_s|^2 + K^2|E_s|^2 + 2(1-K)K|E_s|^2 cos\beta} \\ &\leq (1-K)|E_s| + K|E_s| = |E_s| \end{aligned} \tag{2}$$

The inequality follows the Law of cosines. As long as the $\beta \neq 2k\pi,\ k \in Z$, the signal power $P_s' < P_s$ during propagating.

Regarding the scattering on rough sidewalls, light is decoherent and thus incurs incoherent crosstalk. This noise can accumulate along the topological paths of networks and becomes considerable when it comes to large-scale photonic circuits.

To summarize, due to the imperfectness of the realistic photonic devices, there always exists the phase error. We thereby reinterpret the phase item into,

$$\phi = \phi_0 + \Delta\phi + \Phi_\sigma \tag{3}$$

where ϕ indicates the real phase in consideration of all non-ideality, ϕ_0 connotes the expected phase under ideal conditions, whereas $\Delta\phi$ and Φ_σ are the fixed-value drift and random error, respectively. Note that we assume Φ_σ follows a Gaussian distribution $N(0, \sigma^2)$. They affect the *extinction ratio* (ER).

B. Device-level Crosstalk Model of Unitary Blocks

In the literature of photonic computing, especially for the quantum domain, the port is also defined as *mode*. We continue to adopt the concept in this paper. Hence, the N in $N \times N$ unitary networks denotes the number of modes, namely ports.

We firstly formulate the intra-wavelength crosstalk in unitary blocks. To quantify the derivation of crosstalk, we dissect the MZI-based unitary blocks, as shown in Fig. 2, into three stages. The superscript t denotes the stages of light traverses. The subscripts indicate the indices of modes.

$$\begin{bmatrix} E_{1,\text{ideal}}^{(t+1)} \\ E_{2,\text{ideal}}^{(t+1)} \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} e^{i\Phi} & i \\ ie^{i\Phi} & 1 \end{bmatrix} \begin{bmatrix} E_1^{(t)} \\ E_2^{(t)} \end{bmatrix},\ t = 0, 1 \tag{4}$$

where Φ connotes the phase shifters at the two stages, followed by 3dB couplers. When $t = 0, \Phi = \phi$; when $t = 1, \Phi = 2\theta$.

The essence of loss is the absorption or radiation during propagating or evanescent coupling. The refractive index of realistic materials contains an imaginary term, $\hat{n} = n + i\kappa$. For convenience, we define a symbol L to denote the loss $e^{-C\kappa}$ in the power domain, where $C = \frac{2\pi}{\lambda_0}s$ summarizes all miscellaneous parameters of physical implements. Back to a 2×2 3dB coupler, the light passing and crossing incur the passing loss L_p and the crossing loss L_c, respectively. Remind the power $P = |E|^2$, and the lossy signals are yielded as,

$$\begin{bmatrix} P_1^{(t+1)} \\ P_2^{(t+1)} \end{bmatrix} = \begin{bmatrix} L_p & L_c \\ L_c & L_p \end{bmatrix} \begin{bmatrix} P_{1,\text{ideal}}^{(t+1)} \\ P_{2,\text{ideal}}^{(t+1)} \end{bmatrix},\ t = 0, 1 \tag{5}$$

The lossy field $|E^{(t+1)}| = \sqrt{P^{(t+1)}}$ and keeps the same phase as $E_{\text{ideal}}^{(t+1)}$. In brief, $E^{(t+1)} = \sqrt{P^{(t+1)}}\angle E_{\text{ideal}}^{(t+1)}$ for updating.

Note that the crosstalk consists of coherent crosstalk and incoherent crosstalk. In the worst case, the coherent crosstalk and the signals are opposite in phase, causing the attenuation of signals. To simplify our model, we approximate this part into the loss. Since the incoherent crosstalk can not interact with the original signals, they travel through the networks until they are detected by the photodetectors.

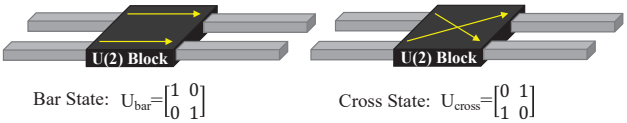


Fig. 3. Bar state and Cross state in a 2×2 Unitary Block. This concept is migrated from switching networks. In the domain of photonic computing, unitary operations are equivalent to fractional combination of Bar and Cross. Note K_1 and K_2 are self-coupling and cross-coupling crosstalk coefficient.

In the switching networks, the beam splitters work in Bar or Cross states. Assume light inputs from Port 1, the crosstalk is $P_{X,2} = K_2' P_{I,1}$ at the Bar state, and $P_{X,1} = K_1' P_{I,1}$ at the Cross state. In MZI-based unitary blocks, the beam splitters are commonly set at 50:50. We migrate this concept from switching, and hence the power of crosstalk can be regarded as the fractional combination. Thus the incoherent crosstalk is,

$$\begin{bmatrix} P_{X,1}^{(t+1)} \\ P_{X,2}^{(t+1)} \end{bmatrix} = \begin{bmatrix} K_1 & K_2 \\ K_2 & K_1 \end{bmatrix} \begin{bmatrix} |E_1^{(t)}|^2 + P_{X,1}^{(t)} \\ |E_2^{(t)}|^2 + P_{X,2}^{(t)} \end{bmatrix},\ t = 0, 1 \tag{6}$$

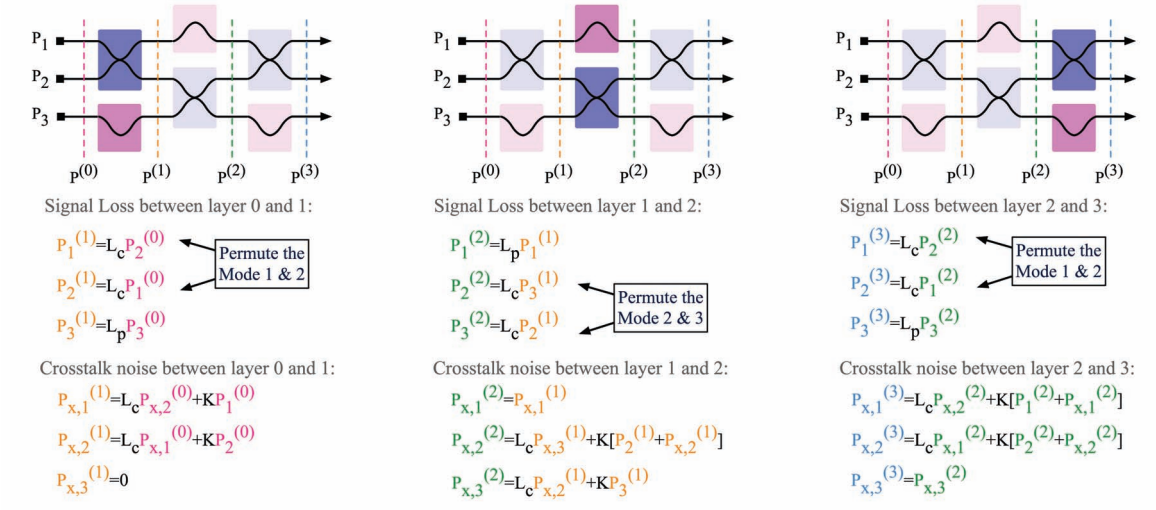


Fig. 4. Illustration of the algorithm for calculating the worst-case loss and crosstalk, layer by layer, for a 3×3 PNN. Rectangles in purple and pink represent unitary blocks, while pink blocks are placed in the margin and only suffer loss. Crosstalk at the initial layer is set to zero. Note that the symbols of passing loss and crossing loss for unitary blocks suppress the subscript u. The calculation of the signal loss and the crosstalk noise follows Equation 11 and 12, respectively. This algorithm also supports symbolic derivation for crosstalk and loss by using *SymPy* libraries in Python. The time complexity is $O(N)$.

C. System-level Crosstalk Model of Photonic Neural Networks

To evaluate the crosstalk in PNNs, depth-first search (DFS) can straightforwardly be applied to find all the adjacent crosstalk paths for targeted output ports. Albeit the DFS-based crosstalk analysis theoretically supports any topological networks, there is no guarantee of the worst-case cost of time since it can be an NP problem. In respect of particular symmetric topologies, the worst-case bounds of crosstalk can be approximately derived. Fortunately, the Clements' scheme, which shapes in rectangular, exists the analytical solution.

An N-mode PNN contains $N(N-1)/2$ unitary blocks in N vertical layers ($N > 2$). To estimate the worst-case crosstalk and loss in PNNs, we extract the loss and crosstalk factors of unitary blocks derived from Equation 4, 5, 6. For conciseness, we denote $L_{u,p}$ and $L_{u,c}$ as the passing loss and crossing loss of unitary blocks, $K_{u,1}$ and $K_{u,2}$ as the self-coupling and cross-coupling coefficients of unitary blocks, respectively. In general, $L_{u,p} < L_{u,c}$ and $K_{u,1} < K_{u,2}$. Under these conditions, given the index of input mode a and the index of output mode b, loss under two extreme cases are,

$$L_{N,min}(a,b) = |a-b|L_{u,c} + (N-|a-b|)L_{u,p} \tag{7}$$

$$L_{N,max}(a,b) \leq \begin{cases} \left[|a-b| + 2f\left(\frac{N-|a-b|-1}{2}\right)\right] L_{u,c} + \\ \left[N-|a-b| - 2f\left(\frac{N-|a-b|-1}{2}\right)\right] L_{u,p}, \\ \quad when: \ N = 2k+1 \\ \\ \left[|a-b| + 2f\left(\frac{N-|a-b|}{2}\right)\right] L_{u,c} + \\ \left[N-|a-b| - 2f\left(\frac{N-|a-b|}{2}\right)\right] L_{u,p}, \\ \quad when: \ N = 2k \end{cases} \tag{8}$$

where a and b are the indices of modes, N connotes the numbers of total modes in the PNN, f represents the floor function. The maximal loss might be less than the right-hand value in Inequality 8 if either a or b hits 1 or N, depending on the implementations of marginal unitary blocks. Note that $N > 2$ and loss is measured in dB unless otherwise specified.

The practical loss should lie somewhere in between,

$$L_{N,\gamma}(a,b) = \gamma L_{N,min}(a,b) + (1-\gamma)L_{N,max}(a,b) \tag{9}$$

where $\gamma \in [0,1]$ summarizes the configurations of PS in paths.

Note that Equation 7 and 8 are derived from topological characteristic but not necessarily matching the constraints of unitary parametrization. According to the properties of unitary matrices, both the loss and crosstalk reach maximum when transforming the anti-diagonal identity matrix [15].

$$U_w(N) = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{10}$$

All unitary blocks in the PNN are supposed to tune π shift, at the Cross state. As aforementioned, since $L_{u,p} < L_{u,c}$, the $U_w(N)$ suggests the worst-case loss and crosstalk.

In this situation, the output mode is the flip of the input mode. For each input-output pair in an N-mode PNN, all the signals from i input mode to $(N-i)$ output mode would traverse 1 passing and $(N-1)$ crossings. Thus the worst-case loss and crosstalk, in ratio (not dB), are,

$$P_{S,\text{worst-case}} = L_{u,p}L_{u,c}^{N-1}P_S \tag{11}$$

$$P_{X,\text{worst-case}}^{(N)} = K[P_S^{(N-1)} + P_X^{(N-1)}] + LP_X^{(N-1)} \tag{12}$$

For the reason that all unitary blocks work at Cross state, only the light passing in a manner of Bar state would be considered as the crosstalk. We thereby simplify the crosstalk coefficients of each unitary block into single K. In Fig. 4, we present a layer-progressive method to calculate the worst-case crosstalk. It expands to series, as shown in Equation 12.

D. Metrics of System Reliability

Prior works quantified the fidelity of the transformation [12] implemented by a lossy N×N experimental interferometer U_e to intended transformation U_i using the following metric:

$$f(U_e, U_i) = \left| \frac{tr(U_i^\dagger U_e)}{\sqrt{N \times tr(U_e^\dagger U_e)}} \right|^2 \tag{13}$$

where tr is the trace of the square matrix, † denotes the conjugate transpose operation.

The fidelity is based on the loss-and-recover model. To put it differently, the fidelity quantifies the ability of schemes to balance the loss among the output modes. For a high fidelity PNN, the signals can be recovered by simply multiplying a post-rectification factor η. However, multiplying such a factor η not only amplifies the signal, but also the noise power. To make matters worse, the variance of noise will expand, that is Var(ηX) = η^2Var(X), which deteriorates the analog-to-digital (ADC) integrity even though we increase the precision depth.

Fidelity obviously fails to summarize the impacts of noise. To address the problem, we use mode-wise signal-noise-ratio (MW-SNR), aimed at estimating the overall signal power of output modes over total noise power. The form of the metric,

$$\text{MW-SNR} = \frac{\sum_i P_{\text{in},i} L_i}{\sum_j P_{X,j}} \tag{14}$$

Since the lossless unitary transformation only changes the hyper-planar angle of input vectors, this property ensures the L2-norm of the optical field is invariant in ideal situations. The numerator and the denominator of MW-SNR represent the overall loss and crosstalk in PNNs, respectively.

E. Analysis of Theoretical Worst-case Bounds

According to our derived Equation 11, 12 and 14, the trends of the worst-case MW-SNR with the increase of the number of modes are shown in Fig. 5. Parameters are shown in Table I.

TABLE I
DEVICE PARAMETERS IN PHOTONIC NEURAL NETWORKS

Parameters	Values
Passing Loss of Unitary Blocks	-0.05 dB
Crossing Loss of Unitary Blocks	-0.10 dB
Propagating Loss of Waveguides	≈ 0 dB
Incoherent Crosstalk Coefficients	$-40 \sim -20$ dB
Power of Input Modes (Uniform)	0 dBm
Post-Rectification Factor η *	11.943

*__Note__: The factor η is obtained by measurement. Configure unitary blocks of the PNN to transform the identity matrix, and input an all-one vector. Then probe the intensity from the output detectors and η is the reciprocal of the average of the intensity.

To manifest the impact of crosstalk at different levels, we conduct a comparison among a set of crosstalk coefficients of photonic devices. Undeniably, the performance of devices, the yields and costs should reach a compromise. For reference, the cutting-edge commercial fabrication can achieve the range of crosstalk coefficients from -30 to -20 dB, probed by ER [16].

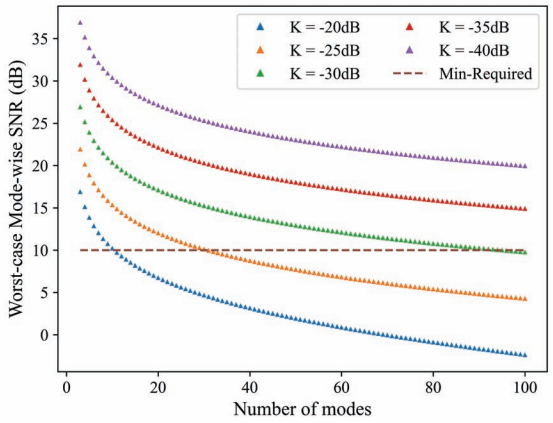


Fig. 5. Trends of the worst-case MW-SNR in various K coefficients. As the number of modes increases, the signal integrity deteriorates faster than what metric fidelity predicts [12]. A PNN monolithic core with crosstalk coefficient $K = -30$dB will encounters the reliability low-bound when mode $N = 96$.

For high-reliability computing, the MW-SNR is supposed to be greater than 20 dB. Even though setting a loose threshold, down to 10 dB, results show PNNs might not be capable of working when $N > 96$ and $ER = 30$ dB. Remind that for an N-mode PNN, the number of photonic device is C_N^2. As opposed to the expectation scaling modes up to 1000 without consideration of crosstalk [17], the number of photonic device integration sharply drop by $C_{1000}^2/C_{96}^2 \approx 109$ times. The results therefore reveal PNNs hardly achieve VLS monolithic integration until the problem of crosstalk is well addressed.

III. IMPACTS ON PHOTONIC NEURAL NETWORKS

In this section, we benchmark the test accuracy of prevailing feed-forward DNN models and compare the performance degradation with and without considering crosstalk and loss. To further investigate the influence of crosstalk on different DNN architectures, we conduct an experiment on multi-layer perceptrons (MLPs) and convolutional layers (CONVs).

A. Experiment Setup

Hardware The hardware specifications of the PNN monolithic core are set as the same as previous in Table I, and K is specified as -30dB. The number of total modes is 120.

Software Modern CNNs adopt the paradigm of CONV-based feature extractors followed by MLP-based classifiers. Since MLPs usually do not have canonical setup of hyper-parameters, we extract the last three layers of the well-known LeNet-5 [18]. These fully-connected layers are equivalent to an MLP architecture. MNIST is selected to be the dataset. As for the experiment on CONV layers, AlexNet [19], VGG-16 [20],

ResNet-18 [21] are selected. We apply the pretrained models from *PyTorch* and modify their classifiers to output 10 neurons, in order to adapt to the CIFAR-10 dataset. Fine-tuning is done by training 5 epochs using SGD optimizer with momentum.

B. Test-Accuracy Degradation of MLP

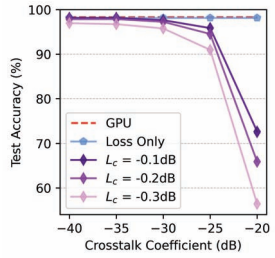
We obtain the input features for the MLP from the outputs of prior convolutional layers of LeNet-5. The convolutional layers are calculating on the electronic GPU and the following MLPs of classifiers are running on 120-mode PNN, shown in Fig. 6.

EXPERIMENTAL SETUP AND PARAMETERS

Layers of LeNet-5	Neurons
Layer 5: Fully Connections	120
Layer 6: Fully Connections	84
Layer 7: Gassian Connections	10
Shapes of Weight Matrices	**Bias**
$M_1 = H \times W = 120 \times 84$	84
$M_2 = H \times W = 84 \times 10$	10

*Note: The matrix M_2 is also processed on the 120-mode PNN core. The rest of unused items are filled with 0.

(a) MLP Configurations



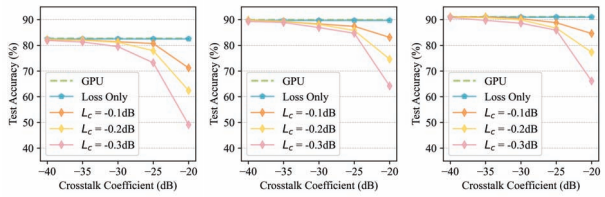
(b) MLP Test-Accuracy

Fig. 6. (a) Configurations and (b) Test-accuracy Degradation of the MLP architecture, with and without considering crosstalk noise. GPU is running on 64-bit precision. Loss-only is in consideration of $L_c = -0.1$dB without crosstalk. The rest of three are L_c with $K = -40, -35, -30, -25, -20$ dB.

Results of MLP evaluations show that test-accuracy continues to degrade as the analogue PNN suffers larger loss and crosstalk. It is worth mentioning that PNN seems robust if only considering loss without crosstalk. But it conflicts with the fact that researchers find signal integrity is bad in VLS-PNNs [22]. And crosstalk model manages to explain the reason and warns us against the severe impacts of crosstalk.

C. Test-Accuracy Degradation of CONV

Before feeding the data into the following CNN models, we transform the shape of images from CIFAR-10 to fit the input shapes $224 \times 224 \times 3$. Convolutional layers are processing on the PNN monolithic core. To improve the computation efficiency, GEMM rearranging and patching technique [23] is applied.



(a) AlexNet (b) VGG-16 (c) ResNet-18

Fig. 7. Test-accuracy Degradation of CNN architectures, with and without considering crosstalk noise. Among three models, the test accuracy of AlexNet drops even more significantly than that of VGG-16 and ResNet-18. According to the trace of crosstalk, it causes by the larger size of the convolution kernels.

In Fig. 7, all three models suffer accuracy degradation as increasing the loss and crosstalk levels. And loss-only model continues failing to characterize the aforementioned facts of performance degradation. Among three convolutional models, the test accuracy of AlexNet drops even more significantly that of VGG-16 and ResNet-18. AlexNet has larger kernel size of 11×11 and signals should propagate at least 11 crosstalk-involved layers before they are probed by detectors. In contrast, most of the layers in VGG-16 and ResNet-18 are stacked 3×3 convolutional layers. Albeit the depth of layers increase, the crosstalk is eliminated because signals encounter detectors as well as A/D, and data are saved in the digital domain instead of maintaining in the analogue domain, indicating that crosstalk is no longer accumulated. Therefore, VGG-16 and ResNet-18 demonstrate the similar robustness over AlexNet in the PNN.

IV. TOWARDS HIGH-RELIABILITY PHOEBE COMPUTING

As the evaluations in the previous section suggest, the reliability of PNN systems can be improved by minimizing the loss and crosstalk, or reducing the number of modes in a monolithic PNN core. In practice, we cannot precisely control or achieve perfectly lossless and non-crosstalk photonic devices. Instead, the number of modes can be dedicatedly selected. According to the evaluation results, intuitively, the architecture of many small-scale photonic cores is a good solution to ensure high reliability while keeping the superiority of time efficiency.

On the other hand, small-scale PNN cores may not able to process the large size of layer weight matrices at a time, so the computation should break down into smaller granularity. This requires more time cycles to proceed. It also needs extra adders to accumulate the sum of all intermediate results and excessive times of A/D conversion together with memory access. In summary, small-scale photonic cores cannot take advantages of ultra-high throughput in analogue photonic computing.

Consequently, the trade-off should lie in between. In this section, we propose a multiple heterogeneous bijou photonic-core architecture (Phoebe) to find out the optimal solution.

A. Phoebe Architecture

Phoebe is shown in Fig. 8. Two layers are over the package substrate. Electronic chiplets are placed at the top layer and photonic chiplets are placed at the bottom layer. Inter-layer chiplets are connected by the bumps linked to on-chiplet wires. Chiplets provide flexible assembly and also improve the yield.

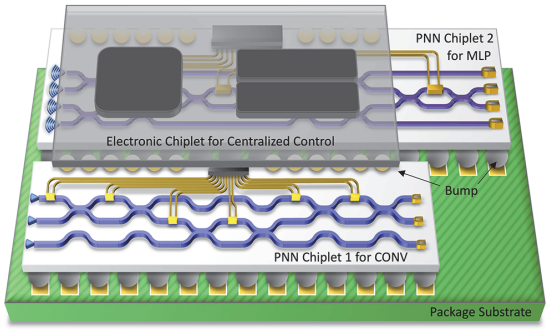


Fig. 8. Overview of Phoebe. It consists of one electronic chiplet at the top layer to schedule the computing tasks as well as memory access, and two heterogeneous photonic chiplets at the bottom layer to accelerate MLP and CONV, respectively. Chiplets between layers are connected by bumps.

B. Superiority of Phoebe Heterogeneous Cores

The mechanism of PNN enables high-efficient computation of fully-connected MLPs naturally. Slicing a big matrix of MLP into several sub-matrices to deploy on more small-scale PNNs, requires more hardware resources, such as detectors. It also introduces complex traffic among small-scale PNN cores. Hence, choosing larger PNN cores but under crosstalk-restraint scale is suitable for MLP-type computation.

By contrast, CONVs need extra preprocessing before computing in PNNs, for instance, patching. It is acknowledged that little kernel sizes such as 3×3, 5×5, 7×7 are more versatile and prevailing in DNN architecture design. This fact suggests that most of the matrices to calculate in PNNs are relatively small matrices with rows of 3^2, 5^2 or 7^2. In this situation, scales of each PNN for CONVs are not necessarily large, or they would suffer considerable crosstalk and loss, leading to insufficient reliability. For that reason, selecting small PNN cores that just fit to kernel sizes is appropriate for CONV-type computation.

Therefore, Phoebe adopts heterogeneous cores on chiplets for type-oriented computation to boost computing throughput and keep high reliability at the same time.

C. Phoebe Benchmark

We conduct an experiment to compare (a) monolithic PNN core, (b) multiple homogeneous cores, and (c) Phoebe in the mixed-type inference tasks. All three have equivalent maximal computing throughput. PNN setup is same as that in Table I. Setup (a) is monolithic 120×120 PNN core. Setup (b) includes 1600 pieces of 3×3 PNN cores. Setup (c) comprises two 60×60, and 50 pieces of 12×12 PNN chiplets.

TABLE II
BENCHMARK OF FULL CNN MIXED TYPE INFERENCE

	Setup (a)		**Setup (b)**		**Phoebe**	
	Time	Acc*	Time	Acc*	Time	Acc*
LeNet-5	0.968	86.5%	1.017	97.8%	1.000	93.2%
AlexNet	0.949	68.1%	1.029	82.2%	1.000	77.3%
VGG-16	0.999	77.3%	1.001	88.4%	1.000	83.2%
ResNet-18	0.987	79.6%	1.013	90.1%	1.000	84.7%

**Time* is normalized cycle counts, *Acc* is test accuracy.
†Four CNNs naturally contain CONV and MLP layers.

It is also worth mentioning that the hardware cost (number of detectors and amplifiers in use) of three setups are 120, 4800 and 720, respectively. As shown in Table II, evaluations show Phoebe achieves high reliability compared with (a), and achieves low cost-per-performance compared with (b). Note that this benchmark has not included the latency of memory access and network traffic. Time of Setup (b) should be longer.

V. SUMMARY AND CONCLUSIONS

To our best knowledge, we are the first time to systematically analyze the impacts of crosstalk and loss in photonic neural networks, and propose a crosstalk-aware model for reliability estimation. Based on this model, we experimentally reveal the severe reliability degradation in analogue photonic computing paradigm with four case studies. To further facilitate VLS photonic integration, we present Phoebe for heterogeneous multi-photonic computing on chiplets. Phoebe takes advantages of massive computing throughput while keeping high reliability.

ACKNOWLEDGMENT

This work is partially supported by Z0546 and InnoHK ACCESS. Authors also thank Shuqing Lin for discussions.

REFERENCES

[1] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[2] S. Chung, M. Nakai, and H. Hashemi, "Low-power thermo-optic silicon modulator for large-scale photonic integrated systems," *Optics express*, vol. 27, no. 9, pp. 13 430–13 459, 2019.

[3] Q. Li, J.-H. Han, C. Ho, S. Takagi, and M. Takenaka, "Ultra-power-efficient 2× 2 si mach-zehnder interferometer optical switch based on iii-v/si hybrid mos phase shifter," *Optics express*, vol. 26, no. 26, pp. 35 003–35 012, 2018.

[4] J. Carolan, C. Harrold, C. Sparrow, E. Martín-López, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh *et al.*, "Universal linear optics," *Science*, vol. 349, no. 6249, pp. 711–716, 2015.

[5] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[6] Y. Wang, X. Li, Z. Jiang, L. Tong, W. Deng, X. Gao, X. Huang, H. Zhou, Y. Yu, L. Ye *et al.*, "Ultrahigh-speed graphene-based optical coherent receiver," *Nature Communications*, vol. 12, no. 1, pp. 1–7, 2021.

[7] J. L. Holi and J.-N. Hwang, "Finite precision error analysis of neural network hardware implementations," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 281–290, 1993.

[8] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *International conference on machine learning*. PMLR, 2015, pp. 1737–1746.

[9] P. Micikevicius *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[10] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, and P. Frossard, "Adaptive quantization for deep neural network," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] Z. Chen, Y. Zheng, F. C. Choong, and M. Je, "A low-power variable-gain amplifier with improved linearity: Analysis and design," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 10, pp. 2176–2185, 2012.

[12] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.

[13] Y. Xie, M. Nikdast, J. Xu, W. Zhang, Q. Li, X. Wu, Y. Ye, X. Wang, and W. Liu, "Crosstalk noise and bit error rate analysis for optical network-on-chip," in *Design Automation Conference*. IEEE, 2010, pp. 657–660.

[14] Y. Shen *et al.*, "Coherent and incoherent crosstalk in wdm optical networks," *Journal of lightwave technology*, vol. 17, no. 5, p. 759, 1999.

[15] B. Yurke, S. L. McCall, and J. R. Klauder, "Su (2) and su (1, 1) interferometers," *Physical Review A*, vol. 33, no. 6, p. 4033, 1986.

[16] C. M. Wilkes, X. Qiang, J. Wang, R. Santagati, S. Paesani, X. Zhou, D. A. Miller, G. D. Marshall, M. G. Thompson, and J. L. O'Brien, "60 db high-extinction auto-configured mach–zehnder interferometer," *Optics letters*, vol. 41, no. 22, pp. 5318–5321, 2016.

[17] S. Pai, B. Bartlett, O. Solgaard, and D. A. Miller, "Matrix optimization on universal unitary photonic devices," *Physical Review Applied*, vol. 11, no. 6, p. 064044, 2019.

[18] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] R. Xu *et al.*, "A survey of approaches for implementing optical neural networks," *Optics & Laser Technology*, vol. 136, p. 106787, 2021.

[23] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Physical Review X*, vol. 9, no. 2, p. 021032, 2019.