This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MDAT.2022.3161599, IEEE Design and Test

On the Impact of Uncertainties in Silicon-Photonic Neural Networks

Sanmitra Banerjee*, Mahdi Nikdast†, and Krishnendu Chakrabarty*
*Department of Electrical and Computer Engineering, Duke University
†Department of Electrical and Computer Engineering, Colorado State University

***Abstract*—Silicon-photonic neural networks (SPNNs) are being explored as post-Moore's law successors to CMOS-based artificial intelligence (AI) accelerators thanks to their ultra-high speed and ultra-low energy consumption. However, their accuracy and energy efficiency can be catastrophically degraded because of the sensitivity of underlying photonic components to fabrication-process variations and run-time uncertainties (e.g., thermal crosstalk). To apply existing component-level uncertainty mitigation techniques to SPNNs, we need to perform criticality assessment to identify susceptible components. In this paper, we explore several fabrication-process variations and run-time uncertainties in optical components and present a hierarchical study of their impact on the MNIST classification accuracy in an SPNN based on Mach–Zehnder Interferometers (MZIs). Simulation results show that the criticality of uncertainties varies significantly based on both the location and the tuned characteristics of the affected components. We also review existing work on techniques to mitigate such adverse impact on SPNN performance.**

I. Introduction

The rise of deep learning as the foundation of most modern artificial intelligence (AI) applications has been fueled by domain-specific AI accelerators that support custom memory hierarchies, variable precision, and optimized matrix multiplication. Modern AI accelerators demonstrate superior energy- and footprint-efficiency compared to GPUs for a variety of inference and some training tasks. With the slowdown of Moore's law, these accelerators approach fundamental limits on their performance due to (i) the limited computational and performance-per-watt capabilities of silicon CMOS, and (ii) the use of low-bandwidth metallic interconnects [1].

Optical computing and communication can potentially overcome both these performance-limiting issues. Computations required in deep learning, such as matrix-vector multiplication, can be performed entirely in the optical domain with high energy efficiency. For instance, with respect to multiply-and-accumulate (MAC) operations, optical computing can achieve a 1000× better energy footprint efficiency compared to the most energy-efficient electronic accelerators today [2]. Additionally, optical interconnects represent a post-Moore's law alternative to replace low-performance metallic interconnects, hence ensuring lower power consumption, higher bandwidth, and lower latency for the communication.

With the advent of silicon photonics, optical components can now be integrated into dense silicon chips using CMOS-compatible manufacturing techniques. Silicon-photonic neural networks (SPNNs) integrate the performance benefits offered by optical computing and interconnects with the low-cost and mature CMOS fabrication process to enable low-latency and energy-efficient optical domain data transport and processing. However, SPNNs are prone to several reliability issues. Imperfections in the optical lithography process lead to variations in critical waveguide dimensions and hence incorrect operation of photonic components. Moreover, mutual thermal crosstalk between adjacent optical components due to convective heat transfer has been observed [3]. These uncertainties, along with the finite encoding precision on tuning parameters, can lead to erroneous matrix-vector multiplication and a consequent loss in SPNN classification accuracy.

In this paper, we present a comprehensive analysis of the impact of uncertainties in SPNNs. In particular, we show that the effect of uncertainties can vary depending on the location and type of affected optical components. The main contributions of this paper are as follows:

- An overview of different uncertainties in SPNNs originating from fabrication-process variations, manufacturing defects, and thermal crosstalk;
- A hierarchical analysis of the impact of different uncertainties on SPNN performance starting from the component level to the system level;
- A framework to identify critical SPNN components where uncertainties can lead to severe performance degradation.

The remainder of the paper is organized as follows. In Section II, we review the fundamentals of SPNNs and the two main classes of SPNNs (i.e., coherent and noncoherent). Section III discusses the various sources of uncertainties in coherent SPNNs and explores prior work on analysis and mitigation of such component imprecision. In Section IV, we present a comprehensive hierarchical study on the impact of random uncertainties in coherent SPNNs. Our analysis can be used to identify the variation-susceptible "critical" components in the network. We draw conclusions in Section V.

II. An Overview of SPNNs

A multi-layer perceptron-based artificial neural network (ANN) maps an input feature vector to an output vector through a series of linear transformations and non-linear activation functions. The neurons in adjacent linear layers (see Fig. 1(a)) are interconnected using weighted edges; these weights are updated during training to change the effect of each input. To mimic this dynamic weighting of connections, silicon-photonic devices can be used to control the optical transmission between

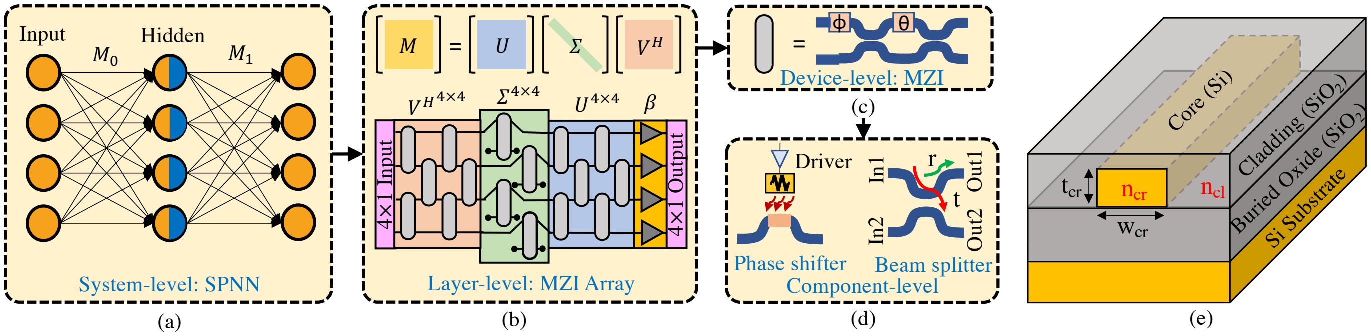


Fig. 1: Hierarchical design of a coherent SPNN: (a) system-level, (b) layer-level, (c) device-level, and (d) component-level; (e) 3D schematic of a strip waveguide.

two neurons in different ways. Coherent SPNNs (C-SPNNs) use thermo-optic phase shifters (PhS) to modify the phase of the optical signal between two neurons. In this case, the tuned phase shifts in the PhS denote the dynamic edge weight. Alternatively, noncoherent SPNNs (N-SPNNs) use microring resonators (MRs) to modify the optical signal power on the interconnection between two neurons. The performance of N-SPNNs can be adversely affected due to geometric variations in the waveguides. Experimental studies have shown that MRs used in N-SPNNs can suffer from a 4.79 nm resonance drift within a wafer due to process variations [4]. Additionally, N-SPNNs require several power-hungry wavelength-conversion steps and are prone to inter-channel crosstalk among different wavelengths.

As a result, C-SPNNs are being preferred for emerging AI accelerators [5]. In this paper, we primarily focus on uncertainties in C-SPNNs. Fully connected layers in C-SPNNs can be represented mathematically as matrix-vector multiplication followed by an activation function. Consider a layer L_i with n_i neurons fully connected to the next layer L_{i+1} with n_{i+1} neurons. The output vector at L_{i+1} is then given by $O_{i+1}^{n_{i+1}\times 1} = f_{i+1}(M_{i+1}^{n_{i+1}\times n_i} O_i^{n_i \times 1})$. Note that f_{i+1} and M_{i+1} are the nonlinear activation function and weight matrix associated with layer L_{i+1}, respectively. In C-SPNNs, the linear multiplication with the weight matrix (i.e., M) is implemented using arrays of configurable Mach–Zehnder interferometers (MZIs), as shown in Fig. 1(b)). Typically, activation functions (e.g., f_{i+1}) are implemented electronically as optical nonlinearities require high signal power and impose lower bounds on the physical footprint [6].

MZIs are used to determine the phase difference between collimated optical signals. Fig. 1(c) shows the typical structure of an MZI with two tunable PhS—with phase shifts ϕ and θ—and two 50:50 beam splitters (BeS). The PhS, shown in Fig. 1(d), are used to apply phase shifts and obtain varying degrees of interference between the optical signals traversing the two waveguides in the MZI. The refractive index of a silicon (Si) waveguide changes with temperature; this is known as the thermo-optic effect. The thermal microheaters in PhS can tune this temperature change by varying the current through a resistor coil. The Joule heat dissipated from the resistor, in turn, controls the applied phase shift. Fig. 1(d) also shows the schematic of a 2×2 directional coupler-based beam splitter. A fraction of the input optical signal denoted by transmittance t in In1 (In2) is coupled to Out2 (Out1) with a phase shift of $\pi/2$. The remaining fraction of the optical signal denoted by reflectance r is reflected to the original waveguide and propagates from In1 (In2) to Out1 (Out2). The ratios r and t are referred to as splitting ratios in this paper. As the optical signal is distributed among the two waveguide in the ratios of r and t, the optical power is distributed in the ratios of r^2 and t^2. Therefore, from the law of conservation of energy, we have $r^2 + t^2 = 1$. In an ideal 50:50 beam splitter, half of the optical power is reflected while the other half is transmitted; therefore, both the transmittance and reflectance coefficients are $\frac{1}{\sqrt{2}}$. The transfer matrix of an MZI with two PhS (ϕ and θ) and two BeS—with splitting ratios (r, t) and (r', t')—is given by:

$$T_{MZI} = \begin{pmatrix} rr'e^{i(\theta+\phi)} - tt'e^{i\phi} & ir'te^{i\theta} + it'r \\ it'r e^{i(\theta+\phi)} + itr'e^{i\phi} & -tt'e^{i\theta} + rr' \end{pmatrix}. \tag{1}$$

Using singular value decomposition, the weight matrix corresponding to the layer L_i can be factorized into two unitary matrices and a diagonal matrix: $M_i = U_i \Sigma_i V_i^H$; U_i and V_i are the unitary matrices and V_i^H denotes the Hermitian transpose of V_i. Moreover, Σ_i is a diagonal matrix consisting of the eigenvalues of M_i. Any $n_i \times n_i$ unitary matrix can be represented by an array of $\binom{n_i}{2}$ MZIs connected as shown in Fig. 1(b). MZIs can also be used to attenuate each waveguide separately without mixing (see $\Sigma^{4\times4}$ in Fig. 1(b)). In this way, an $n_i \times n_i$ diagonal matrix can be represented by n_i MZIs with one input and one output of each MZI terminated using optical waveguide tapers to prevent back-reflection and cross-coupling at the unused ports [7] Additionally, an optical amplification, denoted by β in Fig. 1(b), is required on each output to counter the power dissipation in lossy MZIs.

SPNNs can be trained either in an *in-situ* or an *ex-situ* fashion. In *in-situ* training, gradient computation needs to be performed on the SPNN platform; this involves sequentially perturbing each parameter of the circuit. Such training demands significant computational time and resources and its efficiency can be affected under thermal crosstalk. Thus, current implementations of SPNNs are typically trained *ex-situ* using a software model of the optical system on a digital computer. After training, the voltage drivers in the PhS are configured to realize the trained weights.

III. Uncertainties in SPNNs

Silicon photonic integrated circuits are sensitive to nanometer-scale lithographic variations, manufacturing defects, and thermal crosstalk. In this section, we explore the fabrication-process variations and run-time uncertainties affecting different photonic components.

A. Fabrication-process Variations in C-SPNNs

Imperfections in the optical lithography process may lead to variations in the resist sensitivity, resist thickness, exposure change, and etching. A prominent example of such variation is in the Si waveguide width and thickness. Owing to the high refractive index contrast between the Si core and SiO_2 cladding (see Fig. 1(e)), variations in the waveguide width and thickness significantly perturb the effective index. The effective index (n_{eff}) is the ratio between the phase shift per unit length in a waveguide relative to the phase shift per unit length in vacuum. The effective index also depends on the wavelength of the optical signal.

The temperature-dependent phase shift in PhS is given by $\Delta\phi = \left(\frac{2\pi l}{\lambda_0}\right) \cdot \left(\frac{dn}{dT}\right) \cdot \Delta T$, where l is the length of the phase shifter and λ_0 is the optical wavelength [3]. Also, $\frac{dn}{dT} \approx 1.8 \cdot 10^{-4}\ K^{-1}$ is the thermo-optic coefficient of silicon at λ_0 =1550 nm and temperature T =300 K, and ΔT is the temperature change. The tuned phase shift, $\Delta\phi$ can also change under lithographic variations in l. Additionally, impurities introduced in the waveguide material during fabrication can affect $\frac{dn}{dT}$.

The microheaters in PhS are controlled either by applying a tuned voltage or passing a tuned current across the resistor coil. This voltage/current can be supplied from a DC source based on a digital-to-analog converter (DAC). The precision of the temperature shift ΔT, and in turn, the phase shift is limited by the quantization error in the DAC. For example, in an 8-bit DAC, only 256 different phase shifts in the range [0, 2π] can be realized. Low-precision PhS can degrade the accuracy of the linear multipliers in SPNNs.

The power coupling coefficient in directional-coupler-based BeS denotes the fraction of input power coupled from one member waveguide to the other. This is given by $K(z) = \sin^2(\delta z)$, where z is the coupler length and δ is the field coupling coefficient. In ideal 50:50 BeS ($r = t = \frac{1}{\sqrt{2}}$), $K(z) = 1/2$. Variations in the waveguide dimensions and the gap between the coupled waveguides arising from proximity effects in the etching process affect δ. Changes in δ, in addition to variations in the coupler length z, can lead to non-idealities in BeS.

Fabrication-process variations have a significant impact on the individual PhS and BeS; as a result, MZIs are highly sensitive to manufacturing uncertainties. Indeed, MZIs are more sensitive to differential variations among the two constituent waveguides than the common-mode variations to the entire device. This is because the operation of interferometric devices (e.g., MZIs) depends on the phase difference between optical signals in the constituent waveguides. As a result, common-mode variations, that affect optical signal on both the waveguides uniformly, do not have a significant impact on the performance of MZIs. Clearly, understanding the uncertainties in silicon-photonic circuits (including SPNNs) is essential for yield ramp-up.

B. Run-time Uncertainties in C-SPNNs

Run-time uncertainties in C-SPNNs can arise due to mutual thermal crosstalk among the microheaters in thermo-optic phase shifters. The tuned phase shift in thermo-optic PhS is proportional to $l \cdot \Delta T$, where l and ΔT denote the phase shifter length and the change in temperature, respectively. To minimize the MZI area overhead, larger ΔT is required for tuning PhS. This necessitates increased heater power consumption and results in higher susceptibility to thermal crosstalk. In fact, even the most efficient phase shifter requires a voltage $V_\pi = 4.36$ V and power $P_\pi = 24.77$ mW to provide phase shift of π [8]. The change in phase in the victim phase shifter due to thermal crosstalk depends on its geometric structure, heater material, and the distance from the aggressor phase shifter. For a 5 μm aggressor-victim gap filled with the default SiO_2 cladding and $P_\pi = 24.77$ mW, the optical phase shift in the victim phase shifter is greater than 0.5 rad [3]. Note that due to the latency associated with thermal tuning, the effects of thermal crosstalk may not be localized among proximal microheaters, especially in C-SPNNs with several MZIs. Moreover, due to simultaneous thermal gradients emanating from multiple MZIs, developing a high-fidelity thermal model is complex and requires experimental measurements on a taped-out photonic circuit and is beyond the scope of this paper.

Prolonged voltage biasing of optical components can lead to the formation of traps at the Si-SiO_2 boundary in optical waveguides. Such traps affect the refractive index of the Si core, thereby leading to higher scattering-induced optical loss. Experimental results on on-chip photonic networks show up to a 30% increase in the energy-delay product due to trap-induced aging. Similar aging-induced run-time uncertainties will also affect C-SPNNs due to long-term thermal biasing.

IV. Hierarchical Analysis of the Impact of Uncertainties in C-SPNNs

While there are different sources of uncertainties in PhS and BeS (e.g., lithographic variations, defects, impurities, thermal crosstalk), their impact can be modeled by considering uncertainties in the phase shifts (for PhS) and splitting ratios (r and t for BeS). In this section, we present a case study on the impact of uncertainties in these parameters due to lithographic variations and thermal crosstalk. However, our criticality-assessment approach is agnostic to the source of uncertainties and will therefore hold for any other sources of uncertainties affecting the phase shifts and splitting ratios. Fig. 1(a)-(d) shows the different hierarchical levels in our analysis. Component-level uncertainties in the phase shifters and beam splitters lead to faulty MZI operation at the device-level. An

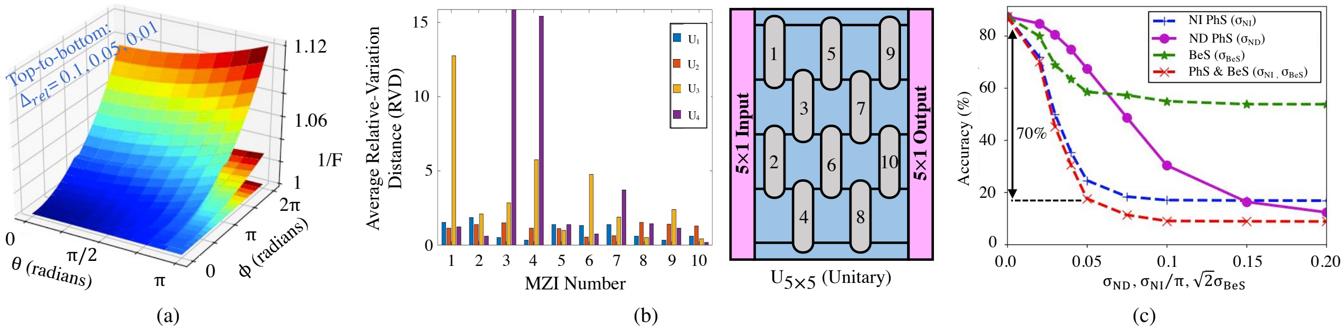


Fig. 2: (a) Deviation in T_{MZI} due to ND phase uncertainties; (b) Average RVD (left) for four 5×5 unitary matrices with one MZI under variations at a time. Right: An MZI array (including the MZI numbers) to represent any 5×5 unitary matrix; (c) Impact of ND and NI phase uncertainties and uncertainties in BeS splitting ratio on the C-SPNN inference accuracy.

array of faulty MZIs lead to deviated matrices (e.g., U, Σ, and V^H); this in turn, leads to faulty weight matrix (M) at the layer-level. At the system-level, a C-SPNN with such faulty weight matrices leads to inferencing errors. We conclude this section with a discussion on few mitigation techniques to improve the tolerance of C-SPNNs against uncertainties.

A. Component-level: PhS and BeS

The phase shifts in thermo-optic PhS can be affected due to lithographic variations in the waveguide, quantization error in the DAC, and thermal crosstalk. Phase uncertainties from these sources can be classified into two main types:

1) *Nominal-dependent (ND) phase uncertainties*: The standard deviation of the phase uncertainties is proportional to the nominal tuned phase shift. In this case, the deviated phase shift is given by $\tilde{\phi} = \phi + \sigma_{nd}\phi\aleph(0,1)$. Here, ϕ and $\aleph(0,1)$ denote the nominal tuned phase shift and the standard normal distribution, respectively. The standard deviation of the uncertainties ($\sigma_{nd}\phi$) increases with ϕ. ND uncertainties predominantly affect PhS with high phase shift; typical sources include thermal crosstalk and quantization errors.
2) *Nominal-independent (NI) phase uncertainties*: In this case, the standard deviation of the uncertainties is independent of the tuned phase shift, $\tilde{\phi} = \phi + \sigma_{ni}\aleph(0,1)$. NI uncertainties include geometric process variations in the waveguide and manufacturing defects and impurities.

Prior studies indicate a mean phase uncertainty of up to 0.21 rad ($\approx 0.07\pi$) in fabricated PhS. To consider a range of uncertainties around this mean, we vary σ_{nd} and σ_{ni} in the range $[0.005\pi, 0.15\pi]$. In ideal 50:50 BeS, $r = t = 1/\sqrt{2}$ (Sec. III.A). However, with uncertainties, a deviation of 1-2% is typically expected in the r and t parameters. For our analysis, we consider the deviated reflectance, $\tilde{r} = r + \sigma_{BeS}\aleph(0,1)$ with the deviated transmittance, $\tilde{t} = \sqrt{(1-\tilde{r}^2)}$. For a fair comparison with the impact of PhS uncertainties, σ_{BeS} is varied in the range $[0.005 \cdot 1/\sqrt{2}, 0.15 \cdot 1/\sqrt{2}]$. Note that uncertainties in the BeS are, in principle, nominal independent as all the devices have the same nominal splitting ratios ($r = t = 1/\sqrt{2}$).

B. Device-level: MZIs

Variations in the phase shifts and splitting ratios affect the MZI transfer matrix, T_{MZI} (1). To measure the closeness between the deviated transfer matrix $\tilde{T}_{MZI}$ and T_{MZI}, we use the fidelity metric given by: $F(T,\tilde{T}) = \left|Trace(\tilde{T}^{\dagger}T)/N\right|^2$. Here, $\tilde{T}^{\dagger}$ and N denote the conjugate transpose and the size of $\tilde{T}$, respectively. Note that $F(T,\tilde{T}) = 1$ if and only if $T = \tilde{T}$ and F decreases with decreasing similarity between T and $\tilde{T}$. Fig. 2(a) shows how F changes due to ND phase uncertainties. In this case, the deviated phase shifts are $\tilde{\theta} = \theta(1+\Delta_{rel})$ and $\tilde{\phi} = \phi(1+\Delta_{rel})$, where Δ_{rel} denotes the relative change in the phase shifts. Clearly, an MZI with higher phase shifts is more susceptible to ND phase uncertainties (the z-axis in Fig. 2(a) denotes $1/F$). However, for NI phase uncertainties, F is independent of θ and ϕ. The susceptibility of different MZIs to such uncertainties, and also to uncertainties in the splitting ratio, depends solely on their position in the MZI array.

C. Layer-level: MZI Array

Unitary multipliers in the linear layers of C-SPNNs can be realized using MZI arrays. Due to faulty MZIs, these unitary multipliers can deviate from their intended form. The deviation can be measured using the relative-variation distance given by: $RVD(\tilde{U},U) = \frac{\sum_{m=1}^{N}\sum_{n=1}^{N}\left|\tilde{U}_{m,n}-U_{m,n}\right|}{\sum_{m=1}^{N}\sum_{n=1}^{N}\left|U_{m,n}\right|}$. Here, $U(\tilde{U})$ denotes the $N \times N$ intended (deviated) unitary matrix and $|U_{m,n}|$ denotes the absolute value of $U_{m,n}$. Fig. 2(b) shows the mean RVD (over 1000 iterations) when uncertainties with $\sigma_{nd} = 0.05$, $\sigma_{ni} = 0.05\pi$, and $\sigma_{BeS} = 0.05/\sqrt{2}$ are inserted in one MZI at a time, in four different randomly generated 5×5 unitary matrices. We observe that the distribution of mean RVD differs across the four unitary matrices. Therefore, the impact of uncertainties in the MZI array on the unitary multipliers depend on both the phase shifts and the position of the affected MZI.

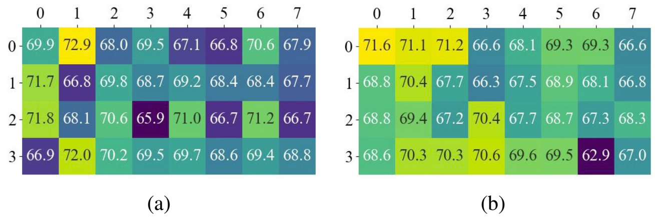
D. System-level: C-SPNN

Incorrect matrix multiplication at the layer-level can lead to misclassifications in the C-SPNN. To understand the impact of uncertainties in the phase shifts and splitting ratios on the classification accuracy, we consider an imprecise fully-connected C-SPNN with two hidden layers of 16 complex-valued neurons. Each linear layer is followed by a nonlinear Softplus layer. A LogSoftMax layer is used after the output layer to obtain a probability distribution. We use a cross-entropy loss function during training. To reduce the feature vector size, each real-valued MNIST image is converted to a complex feature vector of length 16 using fast Fourier transform [9].

Fig. 2(c) shows the mean inference accuracy (over 1000 Monte Carlo iterations) under random ND and NI uncertainties in PhS (characterized by σ_{nd} and σ_{ni}) and uncertainties in BeS (characterized by σ_{BeS}). We observe that for the different cases, the inference accuracy declines steeply due to these uncertainties. In particular, with uncertainties in both PhS and BeS, the accuracy drops by $\approx$70% even under low levels of uncertainties ($\sigma_{ni} = 0.05\pi$ and $\sigma_{BeS} = 0.05/\sqrt{2}$). Also, uncertainties in PhS have a higher impact on the accuracy compared to similar uncertainties in BeS.

Understanding the impact of localized uncertainties in the MZI array is necessary for identifying the critical components in an SPNN. The tolerance of an MZI is defined as the maximum allowable change in the splitting ratio of a component beam splitter that can be recovered using post-fabrication thermal tuning in PhS. Based on this notion of tolerance, it is found that the central MZIs in an array, which require a tuned phase shift very close to 0, have the minimum tolerance to beam splitter fabrication errors. However, the tolerance of an MZI to uncertainties (or the lack thereof) can also be quantified by the accuracy loss due to localized uncertainties in the MZI. A higher accuracy loss signifies lower tolerance of an MZI to localized uncertainties. To simulate the impact of localized nominal-independent uncertainties, we divide the C-SPNN into zones of 4 MZIs (in a 2× 2 grid). We then select one zone at a time to insert uncertainties with $\sigma_{ni} = 0.1\pi$ and $\sigma_{BeS} = 0.1/\sqrt{2}$ while all other zones have background uncertainty with $\sigma_{ni} = 0.05\pi$ and $\sigma_{BeS} = 0.05/\sqrt{2}$. Fig. 3 shows the mean accuracy loss (over 1000 Monte Carlo iterations) due to localized uncertainties in the two unitary matrices corresponding to the first hidden layer in our C-SPNN in the form of heatmaps. Each cell in the heatmaps corresponds to a zone with 2×2 MZIs. The value (color) in each cell denotes the accuracy loss due to uncertainties. We observe that even under similar levels of uncertainties the accuracy loss can vary by up to 10%. Also, note that the low- and high-impact zones are arranged randomly in each heatmap. This reiterates our prior observation that the susceptibility of MZIs to different uncertainties depends on the tuned phase shifts as well as their location in the array.

However, in the presence of nominal-dependent phase uncertainties, the inferencing accuracy is strongly correlated to the tuned phase shift of the affected MZI(s) – MZIs with higher phase shifts are more susceptible to such uncertainties. To demonstrate this, we rank the tuned phase shifts inn each layer of our example C-SPNN in decreasing order, and insert nominal-dependent uncertainties (quantified by σ_{nd}) to the top f_{high}% and bottom f_{low}% ranked phase shifts. Fig. 4 shows the inferencing accuracy loss due to such uncertainties can be catastrophic (up to $\approx$ 60%) when MZIs with higher phase angles are affected. In contrast, MZIs with lower phase angles are practically resilient to ND uncertainties. Therefore, minimizing the tuned phase shifts improves the C-SPNN performance under such uncertainties, in addition to improving their power efficiency (static power consumption in PhS is proportional to the tuned phase shift). However, in realistic scenarios, C-SPNNs encounter both NI and ND uncertainties and therefore the overall susceptibility of MZIs to uncertainties depends on both their tuned phase shift and location.



(a) (b)

Fig. 3: Average accuracy loss (in %) due to zonal perturbations in the unitary weight matrices representing the weights in the first hidden layer: (a) U_{L1} and (b) V_{L1}^{H}.

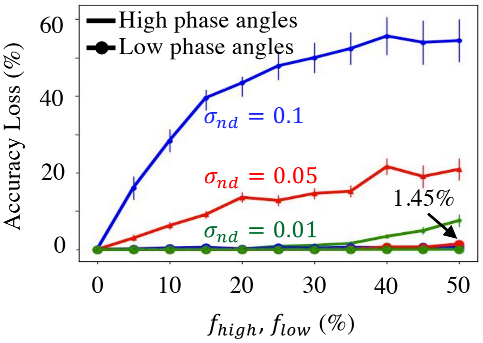


Fig. 4: Comparison between the loss in inferencing accuracy in the presence of ND phase uncertainties in the PhS with the top f_{high}% and bottom f_{low}% phase shifts in each layer.

E. Mitigating the Impact of Uncertainties in C-SPNNs

The extent of the impact of fabrication-process and run-time uncertainties on C-SPNNs has only recently been fully understood and as such, there are very few uncertainty mitigation techniques specific to C-SPNNs. Post-fabrication trimming approaches can minimize the phase uncertainties between the two arms of an MZI by implanting Ge in the Si waveguide. Ge implantation converts crystalline Si (lower refractive index) into its amorphous form (higher refractive index) by breaking the chemical bonds. Due to this, the refractive index (and in turn, the phase shift) in each arm can be precisely trimmed by laser annealing [10]. However, post-fabrication calibration methods rely heavily on the characterization of individual MZIs; therefore, this method is infeasible for C-SPNNs with high MZI count. In order to reduce thermal crosstalk, microheaters can be isolated using deep trenches cutting through the SiO_2 cladding. These structures do not involve special fabrication techniques and lead to a 3× reduction in the phase shift under thermal crosstalk [3]. Recent search efforts for mitigation techniques also focus on uncertainty resilient architectures such as the FFTNet which reduces the optical depth and utilizes fewer MZIs, and the diamond topology where the symmetric structure leads to uniform optical losses in each input-to-output path. An uncertainty-aware training method that uses a modified cost function during training and post-fabrication hardware calibration is presented in [11]. A novel zero-cost optimization technique that improves the power efficiency and robustness by leveraging the non-uniqueness of singular value decomposition has been proposed in [12].

V. Conclusion

SPNNs are prone to nanometer-level fabrication process variations, inter-device thermal crosstalk, optical loss, and manufacturing defects. Each of these sources of uncertainties affects the phase angles and the splitting ratios in different ways. In this paper, we have presented a comprehensive analysis of the various fabrication-process variations and run-time uncertainties and explored several methods to mitigate their impact on the performance of an SPNN. We have used a unified hierarchical approach for criticality assessment of these uncertainties and shown that the degradation in performance depends on both the tuned parameter values and the position of the affected components. Our framework can be used for post-training identification and compensation of critical SPNN components.

Acknowledgement

This work was supported in part by the National Science Foundation (NSF) under grant numbers CCF-1813370, CCF-2006788, and CNS-2046226.

References

[1] D. A. B. Miller, “Device requirements for optical interconnects to silicon chips,” *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1166–1185, 2009.

[2] A. R. Totović *et al.*, “Femtojoule per MAC neuromorphic photonics: An energy and technology roadmap,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 5, pp. 1–15, 2020.

[3] M. Jacques *et al.*, “Optimization of thermo-optic phase-shifter design and mitigation of thermal crosstalk on the soi platform,” *Optics express*, vol. 27, no. 8, pp. 10 456–10 471, 2019.

[4] J. S. Orcutt *et al.*, “Nanophotonic integration in state-of-the-art CMOS foundries,” *Optics Express*, vol. 19, no. 3, pp. 2335–2346, 2011.

[5] M. R. Watts *et al.*, “Adiabatic thermo-optic mach–zehnder switch,” *Optics letters*, vol. 38, no. 5, pp. 733–735, 2013.

[6] I. A. Williamson *et al.*, “Reprogrammable electro-optic nonlinear activation functions for optical neural networks,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–12, 2019.

[7] S. Khan *et al.*, “Low-loss, high-bandwidth fiber-to-chip coupling using capped adiabatic tapered fibers,” *APL Photonics*, vol. 5, no. 5, p. 056101, 2020.

[8] N. C. Harris *et al.*, “Efficient, compact and low loss thermo-optic phase shifter in silicon,” *Optics express*, vol. 22, no. 9, pp. 10 487–10 493, 2014.

[9] S. Banerjee *et al.*, “Modeling silicon-photonic neural networks under uncertainties,” *IEEE/ACM DATE*, 2021.

[10] X. Chen *et al.*, “Post-fabrication phase trimming of mach–zehnder interferometers by laser annealing of germanium implanted waveguides,” *Photonics Research*, vol. 5, no. 6, pp. 578–582, 2017.

[11] Y. Zhu *et al.*, “Countering variations and thermal effects for accurate optical neural networks,” in *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2020, pp. 1–7.

[12] S. Banerjee *et al.*, “Optimizing coherent integrated photonic neural networks under random uncertainties,” to appear in *IEEE/OSA OFC*, 2021.

Sanmitra Banerjee received the B. Tech. degree from the Indian Institute of Technology, Kharagpur, in 2018. He is currently pursuing Ph.D. in Electrical and Computer Engineering from Duke University, Durham, NC. His current research interests include fault modeling and design-for-testability solutions for artificial intelligence accelerators based on emerging technologies. He is a student member of the IEEE and ACM SIGDA.

Mahdi Nikdast is an Assistant Professor in the department of electrical and computer engineering (ECE) at Colorado state university where he leads the electronic-photonic system design laboratory. He received his Ph.D. in ECE from the Hong Kong University of Science and Technology in 2014. He is a senior member of the IEEE.

Krishnendu Chakrabarty received the Ph.D. degree from the University of Michigan, Ann Arbor, in 1995. He is now the John Cocke Distinguished Professor of Electrical and Computer Engineering at Duke University. His current research projects include: design-for-testability of 3D ICs; systolic-array and silicon photonic AI accelerators; microfluidic biochips; hardware security; AI for healthcare; neuromorphic computing systems. He is a Fellow of ACM, IEEE, and AAAS, and a Golden Core Member of the IEEE Computer Society.