OSCAR: An Optical Stochastic Computing AcceleRator for Polynomial Functions

Hassnaa El-Derhalli, Sébastien Le Beux and Sofiène Tahar
Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada
Email: {h_elderh, slebeux, tahar}@ece.concordia.ca

***Abstract*— Approximate computing allows improving design energy efficiency at the cost of computing accuracy. Stochastic computing is an approximate computing technique, where numbers are represented as probabilities using stochastic bit streams. The serial processing of the bit streams leads to reduced hardware complexity but induces high processing latency. Silicon photonics has the potential to overcome this limitation thanks to high propagation speed of signals and high bandwidth. However, the technology remains costly, which calls for optical accelerators capable to adapt to application-specific requirements. In this paper, we propose a reconfigurable optical accelerator capable to adapt to computing accuracy, energy efficiency, and throughput objectives. The architecture can be configured to execute i) 4th order function for high accuracy processing or ii) 2nd order function for high-energy efficiency or high throughput purposes. Evaluations are carried out using image processing Gamma correction application. Compared to a static architecture for which accuracy is defined at design time, the proposed architecture leads to 36.8% energy overhead but increases the range of reachable accuracy by 65%.**

***Index Terms*— nanophotonics, stochastic computing, hardware accelerator, reconfigurable architecture.**

I. INTRODUCTION

Stochastic computing (SC) is an approximate computing technique, in which data are represented as bit streams. Data is processed serially, which contributes to reduce the hardware complexity and the energy consumption [1]. SC is thus suitable for resources-constrained applications that tolerate approximations, such as image processing [2]. A key challenge related to the deployment of the paradigm to a wider range of applications is the high latency induced by the intrinsic serial processing.

Nanophotonics technology is regularly investigated to implement computing architectures, due to signal propagation characteristics, such as low latency and high bandwidth. For instance, the co-integration of photonic and electronic devices on the same die allows implementing microwave processors [3].

Stochastic computing and integrated optics are thus complementary. In [4], we proposed an optical architecture allowing the execution of polynomial functions using the SC paradigm. The architecture relies on a non-linear effect, which allows all optical processing of the data according to the coefficients of the polynomial function. However, the architecture suffers from a limited flexibility to adapt to accuracy requirements.

In this paper, we propose a reconfigurable optical accelerator relying on SC. The architecture can be adapted according to application requirements related to computing accuracy, energy efficiency, and throughput. It can be configured to execute either a 4th order function for high accuracy purposes, or a 2nd order function for high energy efficiency and design throughput. The energy efficiency is estimated using a transmission model we implement, and we evaluate the computing accuracy using Gamma correction application. We carried out a comparison between the proposed reconfigurable and static (i.e., non-reconfigurable) [4] architectures.

II. BACKGROUND AND RELATED WORK

During the last decades, silicon photonics technology has been investigated for the design of optical computing architectures with the aim to accelerate the processing time over electronics-based architectures. Table 1 summarizes related architectures, for which computing complexity ranges from Boolean operations [5] to microwave filters [3]. The computing architectures are implemented using key optical devices we detail in the following.

Table 1: Computing architectures implemented using silicon photonics.

Architecture	Application	Optical devices				Reconfigurability	
		MZI	DC	MRR	AOF	Application	Performance trade-off
Reconfigurable Directed Logic (RDL) [6]	Arithmetic/logic operations			√		√	
Optical LookUp Table (OLUT) [7]	Arithmetic/logic operations			√		√	
Microwave processors [3]	FIR filter	√		√		√	
Optical neural network [8]	Matrix multiplication	√				√	
Optical logic gates [5]	Boolean functions				√		
Optical full adder [9]	Full adder		√				
Optical stochastic computing [4]	Polynomial functions	√		√	√	√	
Reconfigurable optical stochastic computing (this work)	Polynomial functions	√	√	√	√	√	√

A. Silicon Photonics based Computing Architectures

The following introduces computing architectures relying on silicon photonics devices shown in Fig. 1.

***Mach-Zehnder Interferometer (MZI)*:** It is composed of two arms in which the power of an input signal is equally distributed (Fig. 1(a)). On the output side, destructive and constructive interferences are obtained by changing the refractive index in one arm. This device will be used to modulate high power signal. The signal power on the output depends on the Insertion Loss (IL) and the Extinction Ratio (ER), which is estimated using the transmission:

$$T^{MZI}[v] = \begin{cases} IL_{\%}, & v = 0 - \text{constructive state} \\ IL_{\%} \times ER_{\%}, & v = 1 - \text{destructive state} \end{cases} \quad (1)$$

In the context of microwave processors [3], MZI are configured to implement topologies suitable for filters processing (e.g., FIR). More recently, MZIs have been used to implement neural networks [8]. For this purpose, an accurate control of MZIs is carried out to define the losses experienced by signals propagating from a layer to another, hence implementing multipliers.

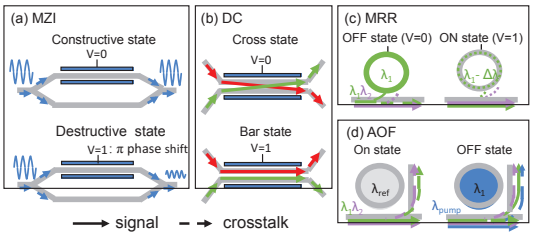

Fig. 1: Silicon photonics devices.

Directional Coupler (DC): Similar to the MZI, a DC is composed of two parallel arms implemented using waveguides (Fig. 1(b)). The device operates in two states: when no voltage is applied, the intrinsic refractive index of the waveguides leads to coupling of the signal from a waveguide to another, i.e., cross state. When a voltage is applied, the change in the refractive index leads to a 50% reduction of the coupling length. Thus, the signals continue propagating on the same waveguide, i.e., bar state. The transmission is defined by:

$$T^{DC}[v] = \begin{cases} IL\_cross_{\%}, & v = 0 \\ IL\_bar_{\%}, & v = 1 \end{cases} \quad (2)$$

In [9], an optical full adder is implemented using DC, which operates as a 2×1 multiplexer. This allows the selection of the right output, i.e., the sum and the carry-out.

MicroRing Resonator (MRR): It is characterized by an initial resonant wavelength λ_i and is controlled using fast electro-optics effect (Fig. 1(c)). It is used to modulate signals as follows: when no voltage is applied (OFF state), the signals at λ_i are coupled into the MRR, which results in a strong attenuation. Whereas, applying a voltage leads to a detuning of the resonant wavelength (ON state), hence the signal transmission to the output is maximized. The transmission φ_t of MRR is given by [10]:

$$\varphi_t(\lambda_{signal}, \lambda_{res}) = \frac{a^2(\lambda_{res})r_2^2 - 2a(\lambda_{res})r_1r_2\cos[\theta(\lambda_{signal},\lambda_{res})] + r_1^2}{1 - 2a(\lambda_{res})r_1r_2\cos[\theta(\lambda_{signal},\lambda_{res})] + [a(\lambda_{res})r_1r_2]^2} \quad (3)$$

where r_1 and r_2 are the self-coupling coefficients, λ_{res} and λ_{signal} are the MRR resonant wavelength and signal wavelength, respectively. $\Delta\lambda$ is the wavelength shift between ON and OFF states, a is the single-pass amplitude transmission, and θ is the single-pass phase shift.

The Reconfigurable Directed Logic (RDL) architecture [6] relies on MRRs organized to implement sum of products. RDL allows executing arithmetic and Boolean functions, such as encoders and adders. The design of Optical LookUp Table (OLUT) [7], using MRRs, takes advantage of Wavelength Division Multiplexing (WDM) to execute multiple functions simultaneously.

All-Optical Filter (AOF): a non-linear effect induced by Two-Photons Absorption (TPA) can be triggered in resonating devices [5]. Indeed, high intensity pump signal allows temporal detuning of the intrinsic resonant wavelength λ_{ref}. In [11], a detuning of 0.1nm for an average 10mW pump signal was reported. As illustrated in Fig. 1(d), signals at wavelength λ_1 and λ_2 continue propagating through the horizontal waveguide when no pump signal is injected. In case a pump signal is injected, the ring resonant wavelength is detuned to a signal wavelength (λ_2 in the example), which leads to the transmission of the corresponding signal to the vertical waveguide. The drop transmission is given by:

$$\varphi_d(\lambda_{signal}, \lambda_{res}) = \frac{a(\lambda_{res})(1-r_1^2)(1-r_2^2)}{1 - 2a(\lambda_{res})r_1r_2\cos[\theta(\lambda_{signal},\lambda_{res})] + [a(\lambda_{res})r_1r_2]^2} \quad (4)$$

In [5], logic gates (e.g., AND, OR and XNOR) operating at 100ps switching time are demonstrated using TPA.

Overall, due to the rather large size of optical devices (typically few μm^2 for MRR [12] to mm^2 for MZI [13]), most architectures are intended to be reconfigurable, i.e., they allow executing a variety of the same type of application. While this configurability level compensates the area and technology complexity overhead, it does not allow adapting the performances, such as throughput, energy efficiency and computing accuracy. Furthermore, the related architectures also suffer from a limited scalability; indeed, the computing paradigms rely on imply a significant number of devices to execute complex applications. For example, the design of a full adder using OLUT [7] requires a total of 23 MRRs. Overall, computing paradigm that intrinsically allows reducing the hardware complexity is needed. SC paradigm offers such a key feature, as detailed in the sequel.

B. Stochastic Computing

In SC, data are represented as stochastic bit streams that are processed serially. This allows computing functions (e.g., addition and multiplication) with limited hardware resources [14]. The computing accuracy depends on bit stream length (BSL) [1]; the longer the stream, the higher the accuracy. SC is commonly used in image processing applications (e.g., edge detection [2] and Gamma correction [15]), neural networks to execute multiplication [16], and signal processing (e.g., FIR filter in [17]).

The Reconfigurable Stochastic Computing (ReSC) architecture [15] allows executing any single input function represented in the form of Bernstein polynomial function given by:

$$B(x) = \sum_{i=0}^{n} b_i B_{i,n}(x) \quad (5)$$

where x is the input data, n is the polynomial order, b_i are the polynomial coefficients, and $B_{i,n}(x)$ is the Bernstein basis polynomial of order n.

As illustrated in Fig. 2(a), the architecture is composed of an adder and a multiplexer. For n order function, n stochastic number generators (SNG) generate the bit streams corresponding to the input data x and n+1 SNGs generate the streams corresponding to the coefficients b_0 to b_n. The bits generated from input x are summed, leading to an n-level control signal. The latter selects the coefficient to be output to a counter. Fig. 2(b) illustrates a processing example for a 3rd order function of x=0.5.

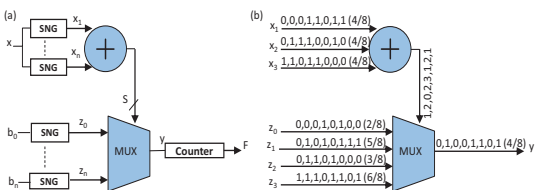

Fig. 2: ReSC architecture proposed in [15].

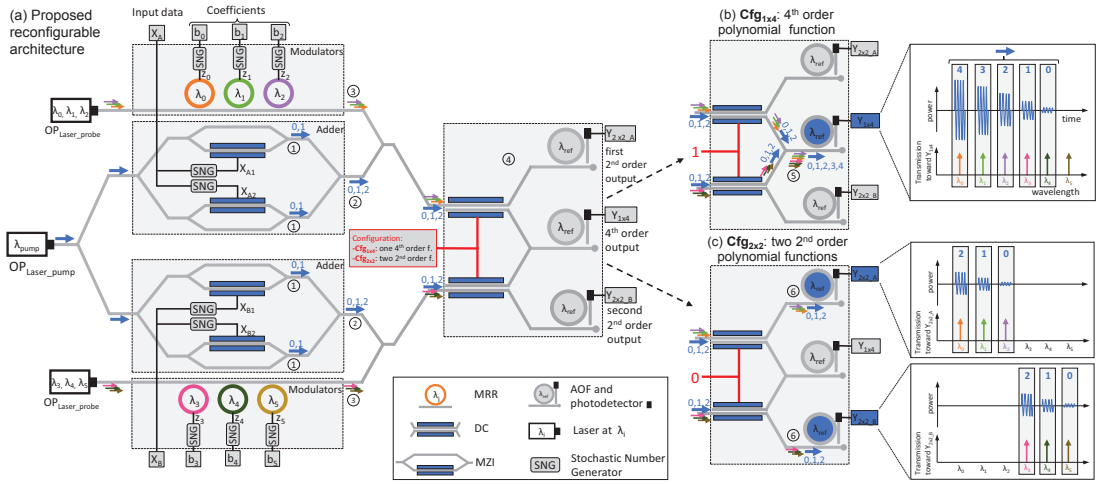

Fig. 3: a) The proposed architecture with the two configurations: b) Cfg_{1x4} leads to a single 4th order function for accurate computing, and c) Cfg_{2x2} leads to two 2nd order functions for high throughput and energy efficiency purposes.

In our prior work, we proposed an optical implementation of the ReSC architecture [4]. MZIs operate as an adder and an AOF implements the multiplexer. While computing accuracy and power consumption are adapted through BSL and laser power, the order of the polynomial function is defined at design time.

In this work, for the first time, we propose an optical accelerator allowing to adapt, at run-time, energy, throughput and computing accuracy. Differently from the static architecture detailed in [4], the proposed architecture allows reconfiguring the polynomial order at run-time. It enables the execution of either a 4th order function (for accuracy purposes) or 2nd order functions (for energy efficiency and high throughput purposes).

III. Proposed Design

In this section, we propose the design of a reconfigurable accelerator and we discuss the design challenges related to the technological and system-level parameters.

A. Reconfigurable Accelerator

Fig. 3(a) illustrates the proposed reconfigurable accelerator. It allows executing polynomial functions on input data X_A and X_B according to Bernstein coefficients (input $b_0..b_2$ and $b_3..b_5$). Two configurations are available: Cfg_{1x4} allows executing a 4th order function on the data (i.e., $X_A=X_B$) and Cfg_{2x2} leads to two 2nd order functions processed in parallel (i.e., $X_A \neq X_B$). Depending on the selected configuration, the results are output either on Y_{1x4} (for Cfg_{1x4}) or Y_{2x2_A} and Y_{2x2_B} (for Cfg_{2x2}).

The reconfigurability involves a symmetrical architecture: two sets of adders and modulators are designed using MZIs and MRRs, respectively. Each one is responsible for generating optical signals corresponding to the related input data (i.e., X_A or X_B) and coefficients ($b_0..b_2$ or $b_3..b_5$). The data signals are generated as follows: from data X_A (resp. X_B), streams of bits X_{A1} and X_{A2} (resp. X_{B1} and X_{B2}) are generated using independent SNGs; their outputs modulate MZIs, thus leading to constructive state (1) or destructive state (0) on signals at λ_{pump} (see mark ① in Fig. 3(a)). Eventually, for each pair of MZIs, three optical power levels can be obtained: 0 for 00, 1 for 01/10 and 2 for 11 (see ②). The optical signals corresponding to coefficients b_i are obtained through modulation of MRRs at λ_i using SNGs, where $0 \leq i \leq 5$ (see ③). WDM allows transporting the coefficient signals simultaneously. The distance between consecutive coefficient signals is defined by wavelength spacing ($WL_{spacing}$). Data and coefficient signals are combined into a waveguide prior entering a reconfigurable multiplexer implemented using DCs and AOFs (see ④). The configuration depends on the states of the DCs, as detailed in the following:

- **Configuration Cfg_{1x4}** involves both DCs in the cross state (Fig. 3(b)). The two groups of data and coefficient signals are combined into the same waveguide as follows (see ⑤): while the coefficient signals combined without interfering due to WDM, data signals cumulate with each other, since they both propagate at λ_{pump}. This leads to five pump power levels able to detune the AOF to five wavelengths at which the coefficient signals propagate. The signal at the wavelength selected by the AOF is dropped to output Y_{1x4}, where the number of '1' is counted for stochastic to binary data conversion purposes. This configuration allows executing a 4th order function.
- **Configuration Cfg_{2x2}** involves both DCs in the bar state (Fig. 3(c)). The two groups of data and coefficient signals continue propagating independently from each other (see ⑥). For each group, the pump signal detunes the corresponding AOF to one of the three wavelengths propagating the coefficient signals (i.e., $\lambda_0..\lambda_2$ for Y_{2x2_A} and $\lambda_3..\lambda_5$ for Y_{2x2_B}). This allows simultaneous execution of two 2nd order functions.

Since DCs enable the switching between a single 4th order function (Cfg_{1x4}) and two 2nd order functions (Cfg_{2x2}), the architecture allows exploring accuracy and throughput tradeoffs at run-time. For image processing applications, the high polynomial order available in Cfg_{1x4} configuration is suitable to meet objectives related to computing accuracy. On the other hand, the parallelism available in Cfg_{2x2} configuration accelerate the

processing, either using data level parallelism (by applying the same filter on multiple images simultaneously) or instruction level parallelism (by applying multiple filters on the same image). However, compared to static architecture [4], this adaptability leads to area and energy overhead. This calls for design optimization with the key challenges introduced in the following.

B. Design Method

The laser powers are key design parameters to optimize. Indeed, while the laser powers should be minimized for energy efficiency purpose, enough optical power should be injected to ensure that the design works properly and the computations are correct. The reconfigurability of the architecture leads to additional constraints, since the same injected pump power should control either a single AOF (Cfg_{1x4}) or two AOF (Cfg_{2x2}). While existing methods allow adapting laser powers at run-time [18], they lead to a significant control overhead we intend to avoid in the context of SC as they would impact both latency and area. Instead, we aim to optimize, at design time, the laser powers taking into account the characteristics of the involved devices, i.e., MZI, MRR, DC and AOF, and system-level parameters, such as BER (Table 2). For this purpose, we investigate the wavelengths of the coefficient signals, since they affect both lasers pump and probe powers.

Table 2: System-level and technological parameters.

	Name	Description	Unit
System	n	Polynomial order	-
	BSL	Bit Stream Length	-
	BER	Bit Error Rate	-
	$WL_{spacing}$	Wavelength spacing between probe signals	nm
MZI	T^{MZI}	Transmission through MZI (Eq. 1)	%
DC	T^{DC}	Transmission through DC (Eq. 2)	%
MRR	λ_i	Resonant wavelength in OFF state	nm
	$\Delta\lambda$	Wavelength shift between ON and OFF	nm
	φ_t	Through transmission (Eq. 3)	%
AOF	λ_{ref}	Resonant wavelength w/o injected carrier	nm
	OTE	Optical Tuning Efficiency	nm/mW
	φ_d	Drop transmission (Eq. 4)	%
Laser	η	Lasing efficiency	%
Photodetector	R	Responsivity	A/W
	i_n	Internal noise current	A

First, we define two groups of wavelengths to be processed in parallel under Cfg_{2x2} configuration. Each group contains consecutive wavelengths, hence the pump power is equally distributed to two AOFs. The total wavelengths range (i.e., from λ_0 to λ_5) is also equally distributed, which allows using the same optical tuning efficiency for all the AOFs. Second, we define for each AOF an initial resonant wavelength λ_{ref} allowing to minimize the covered wavelength distance. For Cfg_{2x2}, λ_{ref} is defined as close as possible to the right-most wavelength in the group (λ_2 and λ_5 for Y_{2x2_A} and Y_{2x2_B}, respectively), which is given by the minimum optical power received by the AOF (i.e., 00), hence it depends on the MZI and DC insertion losses. Finally, a large $WL_{spacing}$ leads to a low crosstalk between the coefficient signals, which minimizes the required lasers probe powers. On the other hand, this requires higher pump power to cover a larger wavelength distance by the AOF. Therefore, the optimal spacing, i.e., the spacing minimizing the total laser power, is searched analytically by exploring the $WL_{spacing}$. The design calls for a transmission model we define in the sequel.

IV. Implementation and Model

The configuration proposed in Section III, allows run-time adaptation of accuracy, energy-efficiency and throughput that comes with power overhead. In this section, we detail the signal transmission model of the reconfigurable accelerator. It allows evaluating the Signal-to-Noise Ratio (SNR), from where the laser energy consumption is estimated. The model is unified and is thus applicable for the two configurations. The configuration is defined by cfg, which controls the state of the DCs (i.e., Cfg_{2x2} and Cfg_{1x4} lead to bar state and cross state, respectively). The coefficient signal λ_i propagates through a) the modulating MRR_i, b) modulators MRR_w dedicated to other signals, c) a DC, and d) an AOF, as defined by:

$$T_{s,z}[i] = \underbrace{\varphi_t(\lambda_i, \lambda_i - \Delta\lambda \times z_i)}_{\text{Modulating MRR transmission}} \times \prod_{\substack{w=j,\\ w\neq i}}^{k} \underbrace{\varphi_t(\lambda_i, \lambda_w - \Delta\lambda \times z_w)}_{\text{Other MRR transmission}} \times \underbrace{T^{DC}[cfg]}_{\text{DC transmission}} \times \underbrace{\varphi_d\left(\lambda_i, \lambda_{ref} - \Delta\text{AOF}(x)\right)}_{\text{AOF transmission}} \quad (6)$$

where s is the optical coefficient signal and z is the value of the coefficient. z_i=1 implies a $\Delta\lambda$ detuning of the MRR (ON state) and z_i=0 leads to an alignment of the modulator with signal at λ_i (OFF state). The attenuation by the other MRRs depends on z_w, while the one experienced in the DC depends on the configuration. Eventually, the transmission on the drop port of the AOF (i.e., to the photodetector) is given by the detuning achieved by the pump signal. The wavelength spacing between consecutive coefficient signals is given by:

$$WL_{spacing} = \lambda_{i+1} - \lambda_i \quad (7)$$

The detuning of the AOF depends on the transmission of the pump signal through the MZIs and the DCs. It is given by:

$$\Delta\text{AOF} = OP_{Laser\_pump} \times OTE \times \frac{1}{n} \sum_{j\in h} T^{MZI}[X_j] \times T^{DC}[cfg] \quad (8.a)$$

$$\begin{cases} h = \{A1, A2\} & , Cfg_{2x2} \\ \quad \{B1, B2\} & \\ h = \{A1, A2, B1, B2\} & , Cfg_{1x4} \end{cases} \quad (8.b)$$

where OTE is the Optical Tuning Efficiency (assumed to be 0.01nm/mW [11]). $T^{MZI}[X_j]$ is the transmission through the MZIs, for which the states (constructive or destructive) depend on the data input X_j. Eq. (8.b) indicates which MZIs will be considered in the transmission according to the selected configuration: either the pump signals are separated (Cfg_{2x2}), or they remain combined (Cfg_{1x4}). The SNR is defined as:

$$SNR = OP_{Laser\_probe} \times \frac{R}{i_n} \times \left(T_{s,z_i=1}[i] - \sum_{\substack{w=0\\ w\neq i}}^{n} T_{s,z_w=1}[w]\right) \quad (9)$$

where R is the photodetector responsivity (1A/W), i_n is the photodetector internal noise (4μA), $T_{s,z_i=1}[i]$ is the transmission of signal i as '1', while the crosstalk signals from other MRRs are transmitted as '0', and $T_{s,z_w=1}[w]$ is the transmission of all crosstalk signals as '1', while signal i is transmitted as '0'. Finally, the Bit Error Rate (BER) is calculated from SNR by assuming ON/OFF keying (OOK) modulation of the coefficient signals.

$$BER = \frac{1}{2} erfc\left(\frac{SNR}{2\sqrt{2}}\right) \quad (10)$$

V. RESULTS

In this section, we evaluate the performances of the proposed reconfigurable accelerator using Gamma correction application. We also evaluate the energy and area overhead compared to a non-reconfigurable version of the architecture.

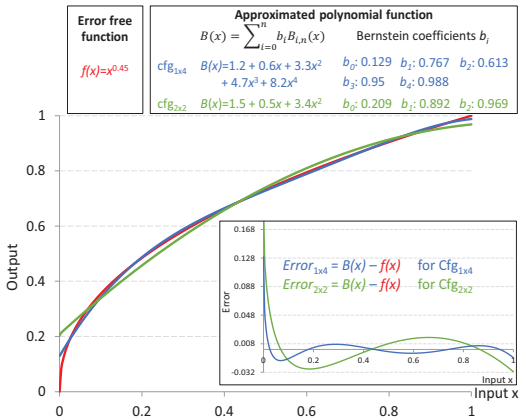

Fig. 4: Error free function f(x) and approximate polynomial functions for Cfg_{1x4} and Cfg_{2x2}.

A. Accuracy and Throughput Trade-off

Gamma correction image processing application [19] is defined as: $f(x) = x^{\gamma}$. We assume γ=0.45, which allows expanding dark pixels into a wider range of values, thus improving the contrast. We aim for an execution on 2nd order (Cfg_{2x2}) and 4th order (Cfg_{1x4}) architectures. For this purpose, the Bernstein coefficients (b_0 to b_2) and (b_0 to b_4) are calculated for Cfg_{2x2} and Cfg_{1x4}, respectively, using the method detailed in [15]. Fig. 4 shows the outputs from processing input data x [0,1] using an error free function $f(x)$, and approximated 2nd and 4th order polynomial functions. As expected, the approximation level increases with the reduced polynomial order, which impacts the error rate and leads to design tradeoff we explore in the following.

To evaluate the architecture, we simulate the processing of 160×160 pixels images for BSL ranging from 2^8 to 2^{12} and BER=10^{-3}. We explore the $WL_{spacing}$, which leads to optimal $WL_{spacing}$=0.155nm. The computing accuracy is calculated using Mean Error Distance (MED), which is obtained by comparing the pixels processed using our architecture with pixels obtained directly from error free results. Cfg_{1x4} leads to sequential processing of the pixels (Fig. 5(a)). For this purpose, each pixel is sent to X_A and X_B and the 5 coefficients are distributed to the MRRs. Cfg_{2x2} is used to process two pixels simultaneously for high throughput purposes (Fig. 5(b)). In this case, X_A and X_B receive different pixels and the same coefficients are sent to the two groups of MRRs. By assuming 1Gbit/s modulation speed and BSL=2^{10}, the average processing time per pixel are 1024ns and 512ns for Cfg_{1x4} and Cfg_{2x2}, respectively. Fig. 5(c) shows the signal transmissions for the two configurations. In the example dedicated to Cfg_{1x4}, we assume a value '1' for the coefficient signals at λ_2 and λ_4 and a value '0' for the remaining λ, thus leading to the transmissions illustrated in ① and ②. The signals are merged and propagate to the same AOF. We assume a received 53mW pump signal power (corresponding to X_{A1}=X_{B2}=1 and X_{A2}=X_{B1}=0), allowing to detune the AOF from λ_{ref} to λ_2 (see ③), thus leading to the transmission of 110µW to Y_{1x4}. For Cfg_{2x2}, we assume the transmission of '1' at λ_2, λ_4, and λ_5 (see ④ and ⑤). The groups of signals propagate to two AOFs, which are detuned independently from each other. The assumed data inputs values lead to the transmission of the signals at λ_1 and λ_5 to Y_{2x2_A}(10µW) and Y_{2x2_B}(90µW), respectively (see ⑥ and ⑦).

B. Static vs Reconfigurable Architectures

Table 3 reports the energy and area overheads of the reconfigurable architecture compared to the static architecture defined in [4]. For a fair comparison, we design our architecture to ensure that Cfg_{1x4} and Cfg_{2x2} achieve the same computing accuracy as the 4th and 2nd order static architectures, respectively. The simulation results show that Cfg_{1x4} and Cfg_{2x2} lead to 53% and 36.8% energy overhead, respectively, which is mainly due to the losses induced by the DCs on the propagation path.

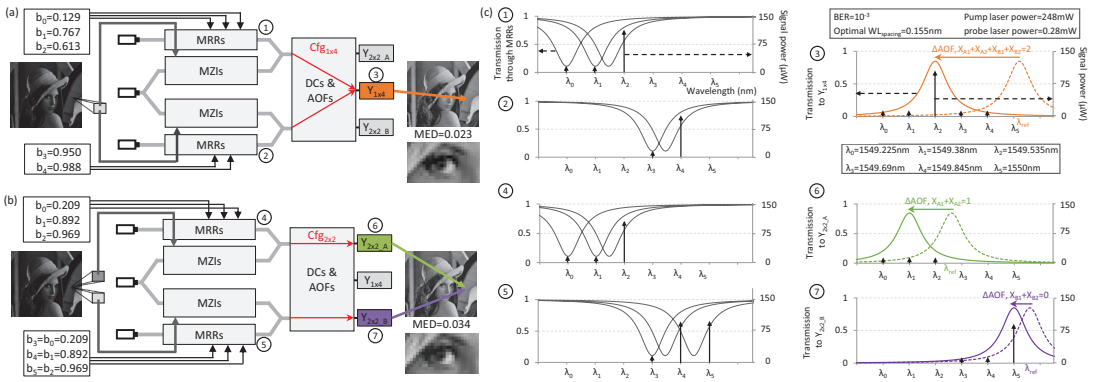

Fig. 5: Image processed for a) Cfg_{1x4}: pixels are serially processed, and b) Cfg_{2x2}: pixels are processed in parallel. ①② and ④⑤ are the transmissions through MRRs for Cfg_{1x4} and Cfg_{2x2}, respectively. ③ and ⑥⑦ are the transmissions towards the photodetectors for Cfg_{1x4} and Cfg_{2x2}, respectively.

Table 3: Energy and area overhead evaluation.

		Static Architecture [4]		Reconfigurable Architecture (this work)		
		n=4	n=2	abs	wrt. n=4	wrt. n=2
Energy efficiency	nJ/pixel	34	19	Cfg_{1x4}:52 Cfg_{2x2}:26	+53% -23.5%	+173% +36.8
Accuracy	MED	0.023	0.034	Cfg_{1x4}:0.023 Cfg_{2X2}:0.034	- +47.8%	-32.4% -
No. of optical devices	Pump laser	1	1	1		
	Probe laser	5	3	6		
	MZI	4	2	4		
	DC	0	0	2		
	MRR	5	3	6		
	AOF	1	1	3		
	Photodetector	1	1	3		
Accuracy/energy adaptability	Order			√		
	BSL	√		√		

We also evaluate the impact of BSL on the computing accuracy and energy efficiency. For this purpose, we evaluate the error and the energy efficiency of all architectures for BSL ranging from 2^8 to 2^{12}. As can be seen in Fig. 6, the proposed architecture allows covering MED ranging from 0.05 to 0.017, while static architectures cover [0.05-0.03] and [0.04-0.017] for 2^{nd} and 4^{th} order, respectively. The improvement in the reachable range of accuracy (+65% and +43.5%) demonstrates the benefits of adapting the polynomial order to satisfy application-level requirements. Interestingly, adapting the polynomial order is, in some cases, more energy efficient than adapting the BSL. For instance, assuming a 2^{nd} order static architecture in Fig. 6, reducing the error from 0.04 to 0.03 can be achieved by increasing the BSL from 2^9 (see ① in the figure) to 2^{12} (see ②), which results in 67nJ/pixel. Using the proposed accelerator, the same computing accuracy can be achieved by switching from Cfg_{2x2} (see ③) to Cfg_{1x4} (see ④), which leads to 26nJ/pixel. It is worth noticing that, in addition to the 61.2% energy saving, a x8 throughput is achieved thanks to a lower BSL (2^9 for ④ wrt. 2^{12} for ②).

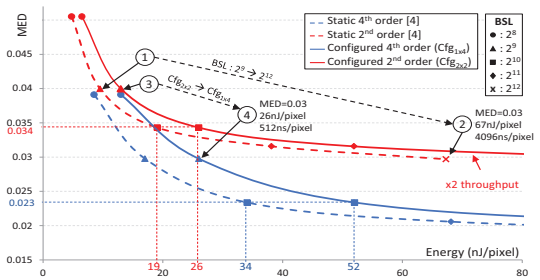

Fig. 6: Accuracy and energy efficiency results to process 160×160 pixels images for BSL ranging from 2^8 to 2^{12}.

To summarize, although the proposed accelerator leads to an area overhead, it covers a large range of computing accuracy, which is needed to adapt to user requirements. This adaptability allows, depending on the targeted accuracy, to improve the energy efficiency or the throughput compared to the static architecture.

VI. CONCLUSION

In this paper, we proposed a reconfigurable optical accelerator relying on stochastic computing paradigm. It allows adapting the order of the executed polynomial functions for accuracy, energy efficiency, and throughput purposes. Compared to a static architecture, in which the order is defined at design time, the reconfigurable accelerator leads to 36.8% energy overhead. However, it increases the range of reachable accuracy by 65%, which is a key to meet users requirements. We also demonstrated that, in some cases, adapting the polynomial order is more energy efficient than adapting the BSL. Future work includes i) the use of power gating to improve the energy efficiency and ii) the design of higher order architectures to further adapt to accuracy requirements.

REFERENCES

[1] A. Alaghi, and J. P. Hayes, "Survey of stochastic computing," ACM Transactions on Embedded Computing Systems, 12(92):1-19, 2013.
[2] R.K. Budhwani, R. Ragavan, and O. Sentieys, "Taking advantage of correlation in stochastic computing," In IEEE International Symposium on Circuits and Systems, pp. 1-4, IEEE, 2017.
[3] D. Pérez, I. Gasulla, and J. Capmany, "Toward programmable microwave photonics processors," Journal of Lightwave Technology," 36(2):519-532, 2018.
[4] H. El-Derhalli, S. Le Beux, and S. Tahar, "Stochastic computing with integrated optics," In Design, Automation and Test in Europe, pp. 1342-1347, IEEE/ACM, 2019.
[5] V. Van et al., "Optical signal processing using nonlinear semiconductor microring resonators," IEEE Journal of Selected Topics in Quantum Electronics, 8(3):705-713, 2002.
[6] Q. Xu, and R. Soref, "Reconfigurable optical directed-logic circuits using microresonator-based optical switches," Optics Express, 19(6): 5244-5259, 2011.
[7] Z. Li, S. Le Beux, C. Monat, X. Letartre, and I. Connor, "Optical look up table," Design, Automation and Test in Europe, pp. 873-876, 2013.
[8] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nature Photonics, 11(7):441-446, 2017.
[9] T. Ishihara, A. Shinya, K. Inoue, K. Nozaki, and M. Notomi, "An integrated optical parallel adder as a first step towards light speed data processing," In International SoC Design Conference, pp.123-124, IEEE, 2016.
[10] H. Li, S. Le Beux, Y. Thonnart, and I. O'Connor, "Complementary communication path for energy efficient on-chip optical interconnects," Design Automation Conference, pp. 1-6, 2015.
[11] V. Van et al., "All-optical nonlinear switching in GaAs-AlGaAs microring resonators," IEEE Photonics Technology Letters, 14(1):74-76, 2002.
[12] W. Bogaerts et al., "Silicon microring resonators," Laser & Photonics Reviews, 6(1), pp.47-73, 2012.
[13] M. Ziebell et al., "40 Gbit/s low-loss silicon optical modulator based on a pipin diode," Optics Express, 20(10):10591-10596, 2012.
[14] B. R. Gaines, "Stochastic computing," Spring Joint Computer Conference, pp. 149-156, ACM, 1967.
[15] W. Qian, X. Li, M.D. Riedel, K. Bazargan, and D.J. Lilja, "An architecture for fault-tolerant computation with stochastic logic," IEEE Transactions on Computers, 60(1):93-105, IEEE, 2011.
[16] R. Hojabr et al., "SkippyNN: An embedded stochastic-computing accelerator for convolutional neural networks," In Design Automation Conference, p. 132, ACM, 2019.
[17] B. Yuan, and Y. Wang, "High-accuracy FIR filter design using stochastic computing," In Computer Society Annual Symposium on VLSI, pp. 128-133, IEEE, 2016.
[18] R. Wu et al., "Variation-aware adaptive tuning for nanophotonic interconnects," International Conference on Computer-Aided Design, pp. 487-493, IEEE, 2015.
[19] R. C. Gonzalez, R. E. Woods, "Digital image processing," Pearson, 2018.