

Review

Thomas Ferreira de Lima, Eli A. Doris*, Simon Bilodeau, Weipeng Zhang, Aashu Jha, Hsuan-Tung Peng, Eric C. Blow, Chaoran Huang, Alexander N. Tait, Bhavin J. Shastri and Paul R. Prucnal

Design automation of photonic resonator weights

<https://doi.org/10.1515/nanoph-2022-0049>

Received January 31, 2022; accepted March 29, 2022;

published online April 11, 2022

Keywords: programmable photonics; RF photonics; silicon photonics.

Abstract: Neuromorphic photonic processors based on resonator weight banks are an emerging candidate technology for enabling modern artificial intelligence (AI) in high speed analog systems. These purpose-built analog devices implement vector multiplications with the physics of resonator devices, offering efficiency, latency, and throughput advantages over equivalent electronic circuits. Along with these advantages, however, often come the difficult challenges of compensation for fabrication variations and environmental disturbances. In this paper, we review sources of variation and disturbances from our experiments, as well as mathematically define quantities that model them. Then, we introduce how the physics of resonators can be exploited to weight and sum multiwavelength signals. Finally, we outline automated design and control methodologies necessary to create practical, manufacturable, and high accuracy/precision resonator weight banks that can withstand operating conditions in the field. This represents a road map for unlocking the potential of resonator weight banks in practical deployment scenarios.

Thomas Ferreira de Lima, Now at NEC Laboratories America, Princeton NJ, USA.

*Corresponding author: **Eli A. Doris**, Department of Electrical and Computer Engineering, Princeton University, Princeton 08544, NJ, USA, E-mail: edoris@princeton.edu

Thomas Ferreira de Lima, Simon Bilodeau, Weipeng Zhang, Aashu Jha, Hsuan-Tung Peng, Eric C. Blow and Paul R. Prucnal, Department of Electrical and Computer Engineering, Princeton University, Princeton 08544, NJ, USA. <https://orcid.org/0000-0002-0362-5183> (T. Ferreira de Lima)

Chaoran Huang, Department of Electrical Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

Alexander N. Tait, Department of Electrical and Computer Engineering, Queen's University, Kingston K7L 3N6, ON, Canada

Bhavin J. Shastri, Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston K7L 3N6, ON, Canada; and Vector Institute of Artificial Intelligence, Toronto M5S 1G1, ON, Canada. <https://orcid.org/0000-0001-5040-8248>

1 Introduction

Most of the recent successes in artificial intelligence (AI) are tied to the ability to perform an increasingly large amount of computation on a mathematical representation of information (i.e., data). In machine learning (ML) algorithms, data is first rearranged into mathematical objects called vectors living inside of a vector space. These vectors can then be manipulated by linear and nonlinear operations resulting in another vector that can be translated into a useful result. For example, in an image classification task the input image is typically encoded into a vector containing the contents of three 2-dimensional arrays of numbers representing the intensity values of red, green, and blue for each pixel. A classifier then computes a function that maps the image vector into a classification vector, which is a 1-dimensional array with each component representing a human-interpretable label (e.g., cat, dog). In modern AI algorithms, especially deep neural networks, most of the hardware processing power in conventional processors is devoted to memory access for parallel linear operations on vectors [1], resulting in a limitation as models scaled up in size. Accordingly, specialized hardware architectures that efficiently perform linear operations have been introduced to accommodate the increasing demand for AI. These include GPUs [2], TPU [3] and neuromorphic processors [4, 5].

In-memory computing and multi-processing techniques are two innovations that have allowed specialized hardware such as multi-core CPUs and GPUs to address the increasing demand for data processing. These techniques, however, were not a significant enough leap forward to close the gap between current computing capabilities and the needs of evolving AI tasks [4, 6], especially for high-frequency computation with low latency. Brain-inspired, or neuromorphic, computing schemes have emerged as a potential alternative for addressing these deficiencies

in power efficiency and speed. Neuromorphic processor architectures are optimized on a hardware level to run neural-network-based algorithms as fast and efficiently as possible [7]. Such architectures derive these benefits from the use of analog systems in which the computation is inherently tied to the physics of the device itself, rather than defined in software on a generalized digital architecture. Analog systems, however, must be engineered to control noise in such a way that the signals of interest are not corrupted by the processing itself, a task which is nearly trivial in digital electronics. It is also crucial to design systems which are user-transparent and are capable of interfacing with existing computational systems.

Most optical phenomena are linear, making multiplication and addition operations easy to implement using photonic devices. Photonic systems can, therefore, benefit from the body of knowledge in the field of ML, which has over the decades created algorithms that are heavily dominated by linear computation. These are directly utilized in photonic processors that operate using analog inputs and outputs. As an added benefit, information in different channels can be encoded onto noninteracting optical carriers using wavelength-division multiplexing (WDM) allowing for a high density of signals on a single waveguide. Utilizing WDM in turn allows for the use of resonator devices that produce controllable effects only around their specific resonant wavelength and do not interact with other optical carriers.

In neuromorphic photonics [8], the device responsible for linear computation is the weight bank, composed of a series of resonators with unique resonant wavelengths, which selectively weights (multiply) a series of incoming WDM signals [9] and optically sums all of the resulting light. Various implementations of resonator weights have been proposed, demonstrated, or leveraged by different groups over the last few years [9–18]. These are optimized specifically for real-time weighted summation of high-bandwidth analog signals encoded onto multi-wavelength lightwaves. Summation using photodetectors returns an analog electrical signal, which can either be read out directly or used to modulate another optical carrier. This multivariate processing ability has proven to be particularly useful in applications with analog inputs and outputs, such as microwave signal processing, ultrafast robotic control [19], and neuromorphic computing [20, 21]. It has also found uses in hardware accelerators for machine learning applications, even with the overhead of converting from digital to analog inputs and outputs [11–14, 17].

One of the main engineering challenges facing weight banks is the ability to precisely and accurately control

weight values, which is made difficult due to fabrication variations and environmental disturbances. In this paper, we aim to outline automated methodologies for designing and controlling resonator weights in order to enable practical WDM photonic processors. In Sections 2 and 3, we introduce the necessary background for understanding and engineering resonator weight banks. Next, in Section 4, we introduce a practical model of weight banks and a feedforward algorithm that can be used for calibration and control of real devices. Finally, in Section 5, we discuss options for designs that allow for compensation of fabrication variations and environmental disturbances as well as a feedback algorithm that can be used if a weight bank is manufactured with built-in weight sensing. The aim is to allow readers to design systems utilizing resonator weight banks that are robust enough to be brought beyond laboratory benches and into the field.

2 Weight bank overview

2.1 Matrix–vector multiplication

Tensor operations, which are the core of many machine learning algorithms, can be implemented via a collection of matrix–vector multiplications. Matrix–vector multiplications, in turn, are typically decomposed into a number of vector–vector dot products where the first vector corresponds to the rows of the matrix and the second vector (the vector of the matrix–vector operation) is identical for all such sub-operations. Hardware units for performing matrix–vector units can be modularized in the same way, with the most basic units implementing vector–vector dot products. Vector–vector dot products can be written mathematically as

$$y(t) = \sum_{i=1}^n w_i(t)x_i(t) \quad (1)$$

where $\mathbf{x} = [x_1, \dots, x_n]^T$ is a vector of incoming signal values, $\mathbf{w} = [w_1, \dots, w_n]^T$ is a vector of weight values, and $y(t)$ is the output. Although $\mathbf{x}(t)$ and $\mathbf{w}(t)$ function identically mathematically, we make the distinction between signals and weights due to the distinct forms that these vectors take in photonic hardware.

Since this operation is a collection of multiplication operations with a summation, it is common to refer to the performance of a system in terms of multiply-accumulate (MAC) operations [22]. The number of MAC operations for a particular computation is hardware-agnostic, although some architectures, such as systolic

arrays [3, 23], implement MACs directly while others, such as the WDM architecture analyzed in this paper, implement weights as individual units and perform all summation simultaneously.

Many applications involve multiplying a batch of vectors with the same matrix, such as running inference tasks with input data on a fixed (pre-trained) neural network classifier. This approach optimizes inference tasks for high bandwidth signals and low latency results. In such cases, it can be advantageous to design analog MAC units where input signals $\mathbf{x}(t)$ can be modulated on fast time scales and are multiplied by “fixed” weights \mathbf{w} that only change on relatively-slow time scales. In general, reconfiguring optoelectronic weights takes significantly longer than their optical signal bandwidth capacity. Optical processors thus need to be codesigned at a system level with an electronic circuitry that optimizes the decomposition of matrix–vector multiplications into smaller vector–vector products while reducing the number of redundant computations [24].

2.2 Role of individual resonators

Resonator-based weights are specifically designed to take advantage of wavelength-division multiplexing (WDM), in which separate signals are encoded on optical carriers with nonoverlapping wavelengths. In optical communications, multigigahertz signals are modulated as amplitude or phase changes on a continuous-wave (CW) carrier traveling through a bus waveguide. With WDM, hundreds of CW carriers with unique wavelengths can coexist in a single waveguide without interfering with one another, effectively creating many independent information channels within a single physical channel. An array of resonators, usually microring resonators, with resonances matching the optical carrier wavelengths is used to access each channel independently. Each resonator, when in resonance, transfers all of the optical energy traveling on one waveguide onto another waveguide. This property is utilized, for example, to enable compact optical interconnections on photonic chips for use in telecommunication systems [25]. If the resonances can be independently tuned, the amount of energy transferred between waveguides can be controlled for each channel, resulting in a precise weighting (multiplication) of the input WDM signal. Using this scheme for implementing matrix–vector multiplications is depicted in Figure 1.

Silicon photonics has emerged as a popular platform for creating photonic integrated circuits (PICs), owing to low loss and compatibility with commercial foundries [26]. Some process design kits (PDKs) are being generated

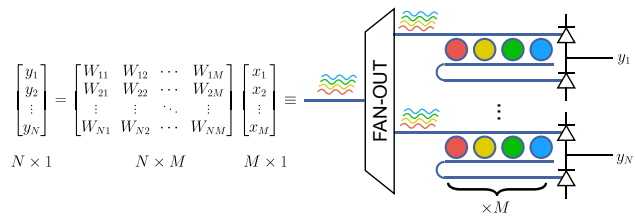


Figure 1: Schematic of WDM-based MVM. Each matrix row of size N corresponds to a weight bank with M resonator weights. At each point in time, the output of the weight bank corresponds to the vector dot product between a matrix row and the multiplicand vector. The higher the bandwidth of the signal $x(t)$, the more computations this system does per second. By stacking N such weight banks, N dot product operations can be completed in parallel without sacrifice to bandwidth.

[27, 28], but optical devices are very sensitive to temperature and fabrication variations. This poses unique challenges to the design of PICs on top of those usually associated with analog devices. In the case of microring resonators, this sensitivity can be counteracted via calibration and control techniques [24, 29]. Early demonstrations used integrated sensors to coarsely lock microring resonators on the transmission wavelength [30]. These involved applying a compensating electronic signal (either current or voltage) based on a measured property, mathematical state model, or both. This can be achieved easily with microcontroller circuits placed in a CMOS electronic circuit nearby or monolithically integrated on the same die [31]. Resonator weights, on the other hand, require more sophisticated control schemes, because the resonator not only needs to be stabilized near resonance, but also need to be fine-tuned with a high degree of accuracy.

3 Resonator weight physics

3.1 Resonator configurations

There are two common configurations for passive resonator weights: all-pass and add–drop, consistent with the terminology used in optical filters. In the all-pass configuration, there is only one waveguide bus coupled to the cavity, with one input port and one ‘thru’ port at the opposite end. On the other hand, in the add–drop configuration, an additional bus waveguide is coupled to the cavity, with one ‘add’ port and ‘drop’ opposite to add. In each configuration, the ports are defined relative to the input port. The ‘thru’ port receives most of the energy from

the input port, except when the cavity is near resonance. Near resonance, a significant amount of the input energy enters the cavity and gets transferred to the ‘drop’ port. For passive, lossless cavities, all energy from the input port is divided between the ‘thru’ and ‘drop’ ports, in a ratio dependent on the input’s wavelength and the coupling between the bus waveguides and the cavity. In theory, the goal is to engineer the resonator so that when on resonance, 100% of the energy drops to the drop port, and when off-resonance, 100% of the energy is transmitted to the ‘thru’ port. In a lossless resonator, this can be simply achieved by designing identical couplers between the all-pass and add–drop waveguides, a condition called *critical coupling*.

In practice, cavities are not perfectly lossless. The main loss mechanisms in silicon photonics are scattering loss (due to roughness of the waveguide’s walls) and material absorption. This loss will perturb the critical coupling condition, causing the power transfer to be unbalanced and resulting in reduced extinction ratio [[32], Sec. 2]. This unbalance can be compensated by design if, and only if, the cavity loss and the coupling coefficients are well known prior to fabrication. These parameters are defined by fabrication and cannot be changed. In a completely passive all-pass microring resonator, for example, this loss is so low that engineering the critical coupling condition is not possible due to fabrication variation.

We can use a simple technique to achieve near-critical coupling, and therefore acceptable extinction ratios, by using a symmetric add–drop bus waveguide, even if we ignore its ports. Because the two couplers are spatially close to one another, the fabrication variation tends to be correlated, ensuring their coupling coefficients are similar. This technique only works if the coupling coefficient is much higher than the cavity loss. In this situation, the intrinsic loss of the microring, dominated by the coupling coefficient to the drop waveguide, automatically matches the coupling coefficient of the input waveguide, resulting in a critical coupling condition. However, as we show in Section 5, sometimes we intentionally modify the cavity’s waveguide via doping or evanescent coupling, increasing the cavity loss to a regime that warrants asymmetric coupling ratios.

The most common example of a photonic resonator is a microring resonator (refer Figure 2). MRRs are standard components in integrated photonics used for filtering, modulation given their simplistic design methodology. They are ring waveguides where the optical path length, i.e., ring circumference, determines the resonant wavelength. By varying the radius, a WDM link with MRRs

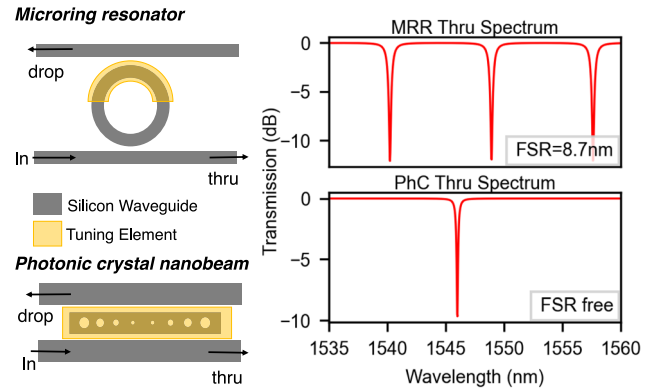


Figure 2: Left: Top-down view of the microring resonator (top), and the photonic crystal nanobeam cavity (bottom) showing the optical ports and the overlaid area where a tuning element can be placed for controlling resonance wavelength. Right: Simulated optical spectra as measured from the thru ports in an add–drop MRR (top) and an add–drop PhC nanobeam (bottom) fabricated on a silicon photonic platform. Note that the PhC cavity has a similar resonance feature but no free spectral range (FSR). The parameters used for the MRR were taken from Table 1, and for the PhC from Ref. [34].

with various resonances can be designed. Besides the resonance wavelength, the ring radius also affects the free-spectral range (FSR) of the cavity, which is the spacing between adjacent resonator modes. FSR ultimately is the limiting factor to the channel count of the WDM link. Sophisticated measures have been employed to avoid this FSR limit, including grating-embedded MRR [33], etc. However, these schemes add to the design complexity and insertion loss. Another way to circumvent this issue would be using photonic crystal (PhC) cavities. Given the relatively strict resonance constraints of PhCs, resonance modes are fewer and can even be engineered to have a single resonance [34]. Eliminating FSR offers advantages in terms of channel count, which may be key for large-scale networks. Nevertheless, the fabrication tolerances of PhC’s require more expensive fabrication techniques such as E-beam lithography which poses a question to its scalable manufacturability.

This article was written with microring resonator cavities in mind, but all the discussions and modeling can be applied to any kind of cavity. We have designed a tool to help engineer MRRs based on known fabrication parameters and waveguide designs [35] which is also detailed in Section 7. Table 1 presents example microring resonator designs useful for weighting, and compares their predicted properties to measured properties. With no fit to experiment, the ring properties can be estimated.

Table 1: Example symmetric add–drop microring resonator designs and their uses.

Device Type	Geometry				Simulated (measured) parameters			Good for
	Width (nm)	Radius (μm)	Gap (nm)	Tuning element	Thru E.R. (dB)	FWHM (nm)	Tuning efficiency	
Weight	500	8	200	n-doped heater	22 (25)	0.4 (0.4)	(0.2) nm/mW	Weighting
Modulator	500	8	300	pn junction	13 (15)	0.1 (0.03)	(0.01) nm/V	High-speed tuning

Waveguide widths are single-valued across bus and ring waveguide. Doping levels are $\sim 10^{17} \text{ cm}^{-3}$ and overlapped with the central portion of the rings away from the couplers, leading to a propagation loss of roughly 5 dB/cm [36]. Parameters were simulated as described in the Appendix without experimental calibration. We compare them to parameters extracted from measurements taken from fabricated rings.

3.2 Context

Signals and Weights: In an analog matrix–vector multiplication, there are two kinds of vectors in a weight bank: the fast vector (\mathbf{x}), which typically represents a signal; and the slow vector, or weight vector (\mathbf{w}), which is understood to be much slower than or stationary compared to \mathbf{x} . In digital hardware, this distinction might seem strange because both vectors are typically loaded from memory at the same rate. In analog hardware, however, it is much more energy efficient to use a slow signal to modulate the amplitude of another signal in a passive way. In photonics, for example, a voltage or current-controlled device can modulate the amplitude or phase of an optical signal, independent of its bandwidth. This is the first caveat for comparing performance between photonic matrix–vector multipliers and electronic ones, especially in terms of energy consumption or speed; in analog computing only one operand can be switched quickly, whereas in digital electronics both weight or signal are functionally equivalent.

One dimension in which optical weights offer a unique advantage is in the latency of the computation. Many applications, such as machine learning, signal processing, and control, require very high bandwidth input signals and low latency but can tolerate slowly-varying weights. For these applications, implementing the MVM in the optical domain is ideal for two main reasons. First, passive optical weights support high bandwidth optical signals with constant (sometimes zero) power dissipation. Second, the latency between input and output is determined by the time-of-flight of the lightwave carrier, or nanoseconds considering resonator and optoelectronic RC-limited bandwidths [37].

Applications such as photonic accelerators, where the computation is offloaded from the CPU to a photonic tensor processor, would not benefit from this scheme since weights would need to be updated as fast as the incoming signals. Doing so imposes a significant energy and speed cost to the control hardware, and is challenging to do

in a way that maintains the high-bandwidth advantage of photonics. However, recent works have been able to marry the highly parallel efficient linearity for nonresonant weight elements, and custom digital electronic ASIC for accelerating deep neural networks [38]. To date, this approach has been favored by industry-oriented start-ups.

Analog Summation: At the output of the weight bank there is a summing element, most commonly implemented with a photodetector or pair of balanced photodetectors (cf. Figure 1). Photodetectors generate a photocurrent proportional to the incident optical power. To first order, they are linear devices and a WDM system acts as a multilinear map, meaning that a linear change in each channel results in a linear change in the output. Realistic photodetectors, however, have nonlinearities at high optical powers, resulting in a saturation curve. In practice, this limits the optical power that can be used in a weight bank. They also typically have nonflat frequency response, so optical signals modulated at different frequencies but with same power generate different levels of photocurrent. Any nonlinear imperfection in the photodetector will contribute to precision and accuracy ‘errors’ in the weighting scheme.

Analog Subtraction: It is often desirable to implement negative weights (in addition to positive weights) in order to allow for the possibility of analog subtraction. Doing so in a WDM system requires the use of add–drop resonators along with balanced detectors, as depicted in Figure 1. In such a case, a channel’s weight value is zero when energy is evenly split between “thru” and “drop” ports.

Mach–Zehnder interferometer (MZI) Approaches: An alternative to WDM-based integrated photonic computational techniques are techniques which utilize Mach–Zehnder interferometers (MZIs) to perform arbitrary linear operations on a single-wavelength optical carrier [39]. Individual phase shifters found in MZIs are typically less sensitive than individual resonators, but process variations compound across large many-MZI PICs in a

way that is not usually seen in large many-resonator PICs. Ultimately, compensating calibration and control techniques are required for practical deployment in either case. A discussion of the most common options for MZI-based PICs can be found in [39].

3.3 Definitions

Normalized Weights: For simplicity's sake, we introduce a normalized weight $\hat{w} \in [0, 1]$. Conversion between real weight w (as measured by the analog summing element) and normalized weight can be accomplished via the equation:

$$\hat{w} = \frac{w - w_{\min}}{w_{\max} - w_{\min}} \quad (2)$$

where w_{\max} and w_{\min} are the maximum and minimum weight values, respectively.

Bit Representation: Before defining error quantities such as accuracy, precision and resolution, sometimes it is useful to refer to them in units of 'bits'. For example, if the relative error of a measurement is 0.125, or 12.5%, we can also refer to it as 3 bits, because it takes three digits to represent that number in binary representation. More generally, if the error is $\varepsilon \in (0, 1]$, the bit representation can be computed as $\varepsilon \text{ (bits)} = \log_2(1/\varepsilon) \in [0, \infty)$. For example, as the measurement error goes to 0, we say that it has infinite precision.

Accuracy: In the context of resonator weight banks, accuracy refers to the systematic error between the commanded weight (denoted $\hat{\mathbf{w}}$) and actual resulting weight vector ($\mathbf{W}(\hat{\mathbf{w}})$), where \mathbf{W} is a random variable for commanded weight $\hat{\mathbf{w}}$. Here, we make the distinction between individual accuracy, which refers to accuracy of any given individual weight, and ensemble accuracy, the average accuracy over all possible individual weights. From a user's perspective ensemble accuracy is a more useful metric, since it reflects expected performance in the general case where weights are unknown ahead of time. Accordingly, we refer to ensemble accuracy wherever it is not explicitly stated. In theory, if laboratory conditions stay stationary over time and instruments are perfectly repeatable, a perfect calibration algorithm would yield infinite ensemble accuracy (zero error). Such an algorithm could visit every possible commanded weight combination, measure the effective weight, and construct a lookup table with the corresponding map. In practice, however, it is desirable to perform this calibration procedure as quickly as possible and with a model that is simple enough to implement in a microcontroller. Ensemble accuracy, therefore, is a metric that measures the quality of the instrumentation, the correctness of the physical model, and the efficiency of the

calibration algorithm. Accuracy is defined mathematically as:

$$\text{Individual accuracy } \delta_{\hat{\mathbf{w}}} \triangleq \|\hat{\mathbf{w}} - \mathbb{E}[\mathbf{W}(\hat{\mathbf{w}})]\| \quad (3)$$

$$\text{Ensemble accuracy } \triangleq \sqrt{\sum_{\mathbf{w} \in \Omega} \delta_{\hat{\mathbf{w}}}^2}, \quad (4)$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbb{E}[\cdot]$ denotes the expected value, and $\sum_{\mathbf{w} \in \Omega}$ denotes the average across all possible weight vectors.

Precision: Precision is complementary to accuracy, and refers to random error caused by noise in the measurement system or control circuit. This is a challenge in any analog system. Here, we again make the distinction between individual precision and ensemble precision, and in general refer to ensemble precision whenever it is not specified. In this case, we chose to construct the ensemble precision so that it includes a contribution from individual accuracies in order to more closely reflect the user's experience. This is intuitively understandable by examining each limiting case. On one extreme, if all individual accuracies are infinite (zero error) then random spread of the ensemble is just the average of individual precisions. On the other extreme, if all individual precisions are infinite (no random spread) then one still expects to see spread in commanded weight values over the entire ensemble due to finite accuracy. Here, we associate ensemble precision to an error bound users would expect from a randomly chosen weight vector. Mathematically, individual and ensemble precision are defined as:

$$\text{Individual precision } \sigma_{\hat{\mathbf{w}}} \triangleq \sqrt{\text{Var}[\mathbf{W}(\hat{\mathbf{w}})]} \quad (5)$$

$$\text{Ensemble precision } \triangleq \sqrt{\sum_{\mathbf{w} \in \Omega} (\sigma_{\hat{\mathbf{w}}}^2 + \delta_{\hat{\mathbf{w}}}^2)}, \quad (6)$$

with the same notation as in the accuracy definition.

Resolution: Resolution ordinarily relates to the discretization error of the signals $x_i(t)$ or $y(t)$, which is not to be confused with weight accuracy and precision. For analog continuous signals, the equivalent terminology is signal-to-noise ratio, or SNR. In this article, we use resolution and SNR interchangeably, often expressed in bits. Because of Eq. (1), the finite precision error of \mathbf{w} causes a degradation in the resolution of $y(t)$. In practice, weight banks are designed such that the resolution is lower than the ensemble precision (e.g., 5-bit precision for a 4-bit resolution), otherwise ensemble precision will determine the resolution of $y(t)$.

Actuation: Actuation refers to the physical process of changing a weight. One common actuation mechanism is

the application of current to a microheater near a resonator. Applied current generates heat, which changes the temperature around the resonator and as a result changes the weight. The actuation mechanism often sets an upper limit on achievable accuracy. For example, consider a resonator controlled by a typical 16-bit DAC within a 0–20 mA range. If the resonator needs 2 mA of current to lock to resonance and 1.5 mA to change the effective weight between -1 and 1 , a single LSB of the DAC corresponds to $0.3 \mu\text{A}$ and $-\log_2(0.3 \mu\text{A}/1.5 \text{ mA}) = 12.3$ bits is the maximum achievable accuracy. In practice, DAC nonlinearities, parasitic resistances, and on-chip thermal crosstalk tend to decrease system accuracy below this upper limit.

Sensing: Sensing is the ability to directly measure each effective weight independently from the actuation mechanism. This plays an important role in simplifying the calibration and control of the weight bank as well as improving accuracy and reconfiguration speed. The ideal sensor would be a spectral sensor that would detect the transmission of each resonator as a function of wavelength and map it to a voltage with high accuracy (0–10 V would be compatible with popular DAC/ADC components currently available from chip manufacturers). More practical sensors measure indirect quantities correlated with the weight, such as local temperature or circulating optical intensity. In this case, the relationship must be incorporated into the calibration and control.

Trimming: Trimming refers to the process of adjusting a physical property of a device to a desired value with high accuracy. This is necessary because common microfabrication processes based on photolithography or e-beam lithography are fundamentally incapable of producing predictable, identical resonator cavities (thus resonance wavelengths) by design. For example, a typical requirement in EDM telecom systems is a fixed operating wavelength; say 1545.32 nm (ITU C-band channel 40). A resonator designed to operate on this channel needs a resonance wavelength accuracy of 0.01 nm, which is currently unattainable by geometric design alone. There are two main ways of achieving trimming post-fabrication. Active trimming involves measuring each resonator's wavelength in advance and using a control circuit to apply a voltage or current bias that compensates for fabrication variations. This approach is referred to as active because it requires the use of external powered circuitry to apply the correction. In contrast, passive trimming involves permanently adjusting the device to the target wavelength via modifications to its physical properties (e.g., phase-change materials or ion implantation). Passive approaches are preferred because they reduce extra control logic and power requirements.

Passive trimming technologies, however, currently require specific post-processing steps that may not be economically viable in silicon photonic foundries.

It is worth noting that active trimming and weighting occur via the same physical mechanisms. The difference between the two, therefore, exists solely in the software layer (i.e., how the user interacts with each). An end user should only need to control weighting, and should not need to worry about active trimming. In other words, the same weight command should produce the same results in two separate chips regardless of manufacturing imperfections. Trimming, therefore, must be completely automatic and free from user interaction. This separation is absolutely critical to enable use of photonic computing hardware by non-experts, and consequently for the proliferation of the technology. In order to realize this, there should be two parallel circuits that are independent and are characterized by independent metrics: trimming, which makes sure that any chip with the same design has the same baseline behavior; and weight control, which translates weight command to effective weight.

4 Control algorithms

One of the biggest advantages of using analog photonic weights is that high-bandwidth optical signals can be weighted by an optoelectronic device that consumes constant power (or static power). It is worth noting, however, that as bandwidth is increased, more static power is required at the optical source to compensate for the added noise captured by the summing photodiodes. In the mechanisms presented in the previous section, the analog weight is defined by the electrical voltage or current applied to the microresonator. The accuracy and precision of each weight is a function of the completeness of the model along with the effects of noise and disturbances. In this section, we outline a simplified weight bank model and discuss practical examples of noise and disturbances.

4.1 Simplified microresonator weight bank model

To illustrate the ramifications of control algorithms, we present a simplified model of a resonator weight bank (Table 2). Each weight (e.g., a microring resonator) can be tuned via silicon's thermo-optic effect by controlling the local temperature. By sourcing current through a resistor placed nearby, this temperature can be adjusted directly and efficiently. This phenomenon is referred to as Joule

Table 2: Simplified microresonator weight bank model, from actuation current (step 1) to resulting weight (step 4).

Step	Physical mechanism
1	<p>Joule heating:</p> $T_i = T_0 + \sum_j^N \mathbf{K}_{ij} R_j I_j^2$ <p>T_i: local temperature; T_0: room temperature; \mathbf{K}: effective thermo-optic coefficient matrix (diagonal in the absence of thermal crosstalk); R_j: electrical heater resistance.</p>
2	<p>Thermo-optic effect:</p> $\lambda_i = \lambda_{0,i} \left(1 + \frac{\beta_{\text{eff},i} \Delta T_i}{n_{\text{eff}}(\lambda_{0,i})} \right)$ <p>$\lambda_i(\lambda_{0,i})$: resonance frequency of filter with applied heat (at room temperature), respectively; β_{eff}: thermo-optic coefficient; $n_{\text{eff}}(\lambda_{0,i})$: effective waveguide index; ΔT: temperature offset.</p>
3	<p>Lorentzian transfer function:</p> $\mathcal{L}(\lambda; \lambda_0, \Delta\lambda) = 1 - (1 + (\lambda - \lambda_0)^2 / \Delta\lambda^2)^{-1}$ <p>$\Delta\lambda$: half-width half-maximum of Lorentzian transfer function modeling a single resonator weight.</p>
4	<p>Weighted sum:</p> $\mu = \frac{1}{N} \sum_i^N \mathcal{L}(\lambda_{l,i}; \lambda_{0,i}, \Delta\lambda_i) \cdot P_i / P_{\text{max}}$ <p>$\mu \in [0, 1]$: weighted sum (photocurrent normalized by PD's responsivity); P_i: optical intensity of input lightwave at wavelength $\lambda_{l,i}$; P_{max}: maximum optical intensity.</p>

Changed to table formatting in response to reviewer 1.

heating. We assume, for simplicity, that the properties of the resistor do not change with its temperature or with the circulating optical power in the resonator cavity. Due to the phase-matching condition on resonators, there is a direct relationship between resonant wavelength and refractive index, namely $\lambda/n = \text{constant}$. In a simple model, the transmission of the microresonator is a function of only the “detuning,” or the difference between the fixed wavelength (λ_0) and the tunable resonance frequency (λ). In high Q -factor resonators, this function can be approximated by a Lorentzian line shape. In Table 2, this is denoted by $\mathcal{L}(\lambda - \lambda_0)$. The second-order deviations from this model are: dependence of transmission on optical intensity (at λ_0), optical crosstalk from neighboring resonators, and undesired optical attenuation in the presence of electrical carriers (most important for doped waveguides).

After weighting, the output of the weight bank is fed to a photodetector. In our simplified model, this photodetector is assumed to have a constant responsivity independent of input power. The electrical properties of the photodetector (parasitic capacitance, parasitic resistance, or frequency response) can be ignored at this point and

instead captured later by the model of the electrical receiver (in neuromorphic photonics, this would be the E/O conversion of the neuron). In practice, the photodetector's responsivity can have a range of values depending on fabrication variation and operating wavelength, and this should be captured by the calibration model via a linear correction. Once the optical power input is too high, however, the responsivity drops due to space-charge screening effects [40]. This nonlinear relationship between photocurrent and optical intensity cannot be calibrated away, and results in reduced precision.

4.2 Compensating for noise and disturbances

The simplified model presented in Section 4.1 is effective at modeling the weighted sum μ as a function of the optical intensities at each wavelength P_i . In this scheme, the signal is modulated onto each carrier wavelength. To simplify, we assume a perfect modulation scheme where the instantaneous intensity is described as follows:

$$P_i(t) = P_{i,\text{input}} x_i(t), \quad (7)$$

where $x_i(t)$ is the same as in Eq. (1) and $P_{i,\text{input}}$.

In this analog representation, we can identify $\mu(t)/P_0$ as the equivalent of the weighted sum $y(t)$ scalar, and the weight as:

$$w_i \equiv \mathcal{L}(\lambda_{l,i}; \lambda_{0,i}, \Delta\lambda_i) \cdot P_{i,\text{input}} / NP_0, \quad (8)$$

where P_0 is a normalization constant.

As revealed by Eq. (8), however, noise and temporal disturbances to the values of $P_{i,\text{input}}$ can negatively affect the weight's precision or the resolution of the sum y . These effects cannot be captured by any calibration model, since by definition a calibration model must be based on static parameters. Disturbances must nonetheless be corrected independently from the transmission control, since the user will reasonably expect the chip to have a constant average weight vector.

In practice, there are two relevant sources of disturbances to the input intensity. One is ambient temperature, and the other is polarization drift accumulated in the input fibers. To show the impact these sources have on the operation of a weight bank, we have setup a microring resonator weight bank with four rings, placed on top of a temperature controlled base (Figure 3). These effects can be observed, for instance, in the optical spectrum of a microring resonator weight bank (Figure 4). We observe that the temperature drift essentially maintains the shape of the transmission curve for each resonator, as expected

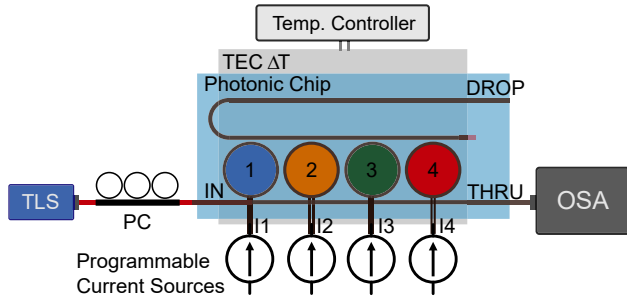


Figure 3: Experimental setup of an optical transmission spectrum measurement of a 4-microring resonator weight bank. The PIC sits atop a thermoelectric module whose heat flow can be electrically controlled, and acts as a “thermal bath” that stabilizes the PIC temperature. A tunable laser source (TLS) generates continuous-wave laser lightwave whose wavelength is quickly scanned. An optical spectrum analyzer (OSA) synchronized with the TLS collects snapshots of the transmission spectrum. A polarization controller (PC) is used to emulate large polarization drifts due to long-term environmental fluctuations.

for microring resonators. As a result, the microcontroller can compensate for that by either dissipating extra heat uniformly across all devices, or by using a thermal bath that holds the average temperature of the entire chip constant, such as the thermoelectric module sketched in Figure 3. This invariance property is unique to microrings and cannot be generalized to other kinds of resonator structures. Keeping temperature constant, however, will stabilize any temperature-dependent device such as resonators and optical couplers. A polarization drift in the fibers outside the chip, however, alters the power spectrum irregularly, and can only be remedied by recalibrating the circuit. We have found that using polarization-maintaining fibers and optical adhesives can stabilize the polarization spectrum indefinitely in laboratory settings. More studies must be completed in order to evaluate this effect in field conditions, where significant contributions are expected from mechanical stress and vibrations.

4.3 Measuring effective weights

The summing element is an integral part of the photonic circuit, so we typically only have access to output photocurrent while operating the weight bank. This creates the challenge of independently verifying that the applied weight was, in fact, the correct one. Mathematically speaking, in the linear approximation of the summing element, for a set of signals $x_i(t)$, and a measurement of $y(t)$, we need to recover the weights w_i in Eq. (1). One approach could be to ensure that only one channel has

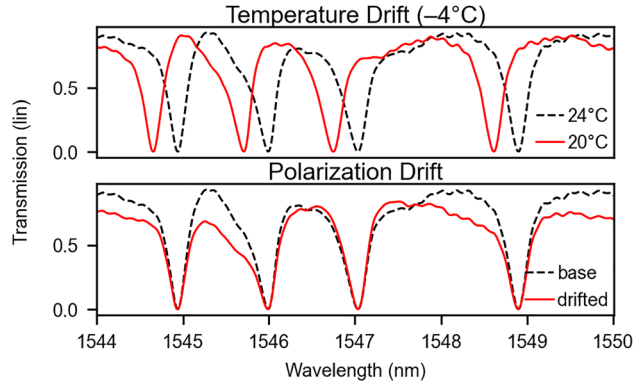


Figure 4: Experimental results of a weight bank’s transmission profile when subject to a temperature drift (top) or an input polarization drift (bottom). When the stage was cooled from 24 °C to 20 °C, the resonance wavelengths of each ring blue-shifted, as expected from the positive coefficient of the thermo-optic effect in silicon. The polarization drift was induced by rotating the planar polarization angle between the laser and the input coupler (cf. 3), resulting in imperfect coupling to the chip and unpredictable changes to the transmission spectrum.

a positive value at any given time, for example $x_i(t) = 1$ for time slot $t \in (i\Delta t, (i+1)\Delta t]$ and $x_i(t) = 0$ otherwise. Then, the measurement of $y(t)$ yields $y(i\Delta t) = w_i$. A better approach, which is more often used, is to use independent (orthogonal) signals, e.g., $x_i(t) = \sqrt{2} \sin(\omega_i t)$ which have the property that $\int x_i(t)x_j(t) dt = \delta_{ij}$. Then, a single measurement of $y(t)$ can be used to compute all weights with the following decomposition: $w_i = \int y(t)x_i(t) dt$. With this method, n weights can be continuously monitored for every integration step.

4.4 Calibration and feedforward control

Regardless of the weighting mechanism, the control strategy of a resonator involves a calibration stage and a control stage. The calibration is required to adjust the weight bank model parameters caused by fabrication variation or uncertain initial conditions. The control stage runs continuously thereafter.

Calibration and control strategies must take into consideration an array of resonators in the weight bank, because coupled resonators suffer from both optical and thermal/electrical crosstalk. There is no noticeable electrical crosstalk in this device, since all heaters are surrounded by silicon dioxide, a highly insulating material. As shown in Figure 5, applying Joule heating to the vicinity of one resonator directly changes its transmission function, which according to the model introduced in Section 4.1 results in changing the weight for that wavelength. Notice,

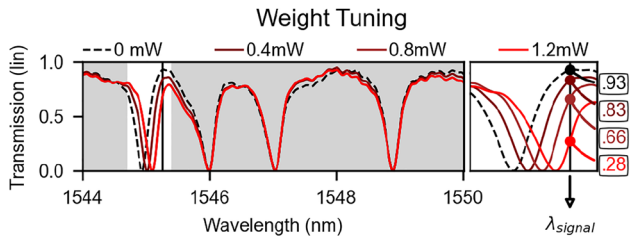


Figure 5: Experimental optical spectral response of a four-ring weight bank as current flows through the heater situated on top of the first ring in the bank (cf. 3). To first order, only one resonance feature is shifted, without affecting others. The inset on the right corresponds to the unshaded area of the left, and shows the transmission values for each of the four tuning powers shown in the legend above. Note that the resonance shift is approximately proportional to the electrical power dissipated on the heater element.

however, that the neighboring resonator also undergoes a small change in its transmission function due to optical crosstalk. Though barely visible, this second order effect, if uncompensated, reduces the accuracy of the weight bank. For a 8-bit accuracy level, the transmission on the other rings must be kept within 0.4% of their target. In our model, we can capture this crosstalk in the matrix \mathbf{K} , shown in Table 2.

Crosstalk can be significantly, but not totally, reduced by carefully engineering the geometry of each device. To reduce the thermal crosstalk, MRRs can be spaced further apart, taking advantage of the low thermal conductivity of silicon dioxide cladding [29]. Optical crosstalk can be reduced by increasing the wavelength spacing between resonators, at the expense of spectral capacity [9].

As the number of weights N in a system increase, so does the number of independent variables in the model. Mitigating the thermal and optical crosstalk between resonators in a weight bank requires measuring crosstalk terms and fitting a multidimensional model with $\mathcal{O}(N^2)$ variables. For example, the crosstalk matrix \mathbf{K} is of dimension $N \times N$. Consequently, we must rely on indirect measurements and physical models to control each resonator in the weight bank system.

A series of approaches have been developed to tackle the weight bank control problem. The earliest method presented in the literature is called the feedforward scheme [41], shown in Figure 6. In this scheme, the controller would possess an electrical model of the resonators, including an approximation of the thermal and optical crosstalk effect. The model maps a vector of electrical current values to a vector of weight values, and vice versa. Based on a desired weight, the model converts to

an actuation electrical signal for each resonator. But once signals are set, there is no way to verify or validate that the effective weight is correct.

The advantage of this approach is the simplicity of the control algorithm – it is similar to a lookup table in electronics. This calibration algorithm scales with $\mathcal{O}(N)$, despite the fact that the feedforward model has $\mathcal{O}(N^2)$ parameters. However, this approach has two main disadvantages: the calibration step is complex and does not work when the outputs of the weight bank are not accessible. It also fails to correct for environmental variations or laboratory conditions. As a result, the chip needs to be recalibrated before each use, making this approach impractical for large N .

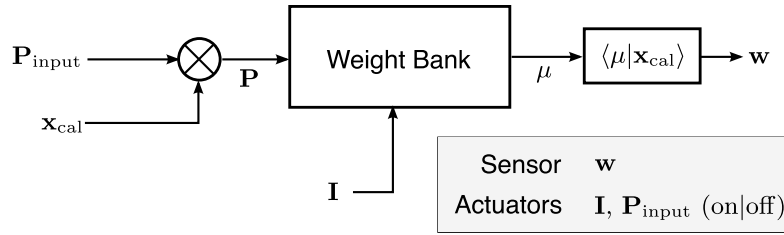
To resolve these fundamental issues with the feedforward scheme, the microresonator device can be designed with an embedded sensor, capable of measuring the applied weight in real time. With a sensor located near each resonator in the weight bank, measuring, e.g., the local temperature, we can directly feed an electric signal back to the controller to adjust for model deviations. This is called a feedback scheme, and will be revisited in Section 5.3, after we delve into index modulation, actuation and sensing physics in photonic waveguides in Section 5.

5 Design for manufacturability

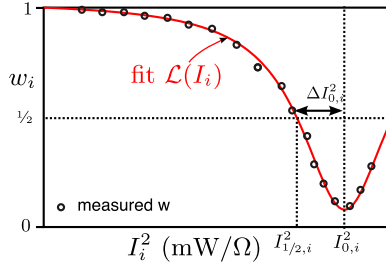
Passive optical devices are sensitive to the effective refractive index of waveguides, which determines the optical path length along the path of propagation. In integrated photonics platforms, the effective refractive index varies with the height and width of the waveguide [42]. The width of integrated waveguides is lower-bounded by a photolithography process with a limited resolution. In the case of silicon photonic chips, deep ultraviolet photolithography is used, with a wavelength of 193 nm, limiting the minimum lateral feature size of devices to 65 nm [26]. Wider waveguides are typically used, however, since increasing width leads to tighter optical mode confinement and lower losses. This, in turn, is upper-bounded by the width at which the waveguide supports a second lateral mode at the wavelength of operation. As a result waveguide width is usually chosen to be the maximum single-mode width minus some engineering margin. For similar reasons, the height of the waveguide is often chosen to be the maximum that only admits a single infinite-slab mode minus some engineering margin. A *de facto* industry standard height is 220 nm with a standard deviation of 2 nm [42].

The combination of lateral and vertical waveguide manufacturing imperfections results in fixed, random

(a) Calibration Procedure



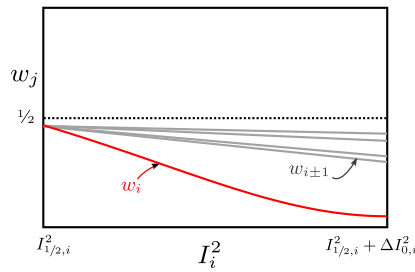
(b) Step 1



For each weight i , set P_i on and sweep I_i . Set all P on. For each weight i , sweep I_i and measure w_j .

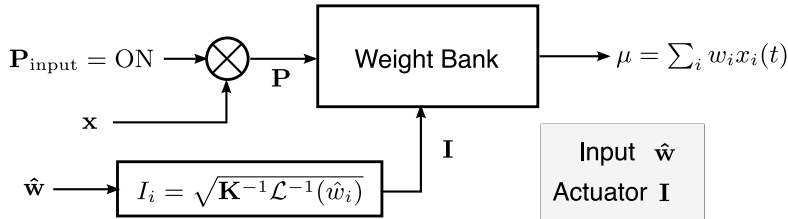
Fit curve and get K_{ii} , $I_{0,i}^2$ and $\Delta I_{0,i}^2$.

(c) Step 2



Fit lines for each $w_j \forall j \neq i$ and get K_{ji} .

(d) Control Procedure



phase shifts in each waveguide both across the same chip and between chips. That means that identically-designed interferometric devices will behave differently once fabricated. For resonator structures, random phase offsets above 2π are sufficient to render any prediction of a post-fabrication resonance wavelength impossible. However, within the same chips the phase offsets are spatially correlated [43], meaning that resonators in close proximity have resonance offsets similar in sign and magnitude.

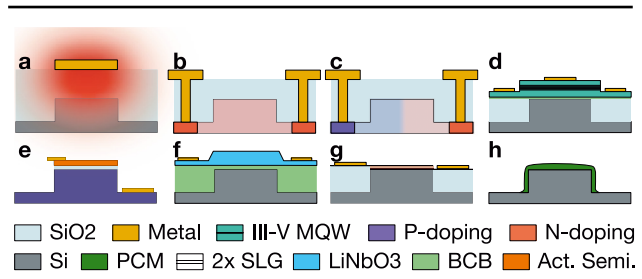
There are several post-fabrication options available to compensate for undesired random phase shifts. These techniques are referred to as trimming, where a section of an optical waveguide material is physically or chemically altered until the desired phase is achieved. The range of phase shifts that can be compensated is typically limited, and as a result design-level compensation is often required. Design for manufacturability tackles this challenge by

predicting the amount of compensation required via Monte Carlo simulations of fabrication variability [44]. A successful design for manufacturability is able to reduce the range of fabrication variation that will need to be compensated by trimming, and can result in considerable cost saving overall.

5.1 Index tuning mechanisms

Many refractive index tuning mechanisms have been developed for silicon photonics, and Table 3 displays some state-of-the-art devices from the literature. In silicon, the strongest effects are the thermo-optic effect, free-carrier absorption and free-carrier dispersion (also known as plasma dispersion). Exploiting the thermo-optic effect for tuning with metal filament microheaters is the easiest and most popular way to effect large index changes, but it is slow and power inefficient [45]. Thermal tuning

Figure 6: Procedure for feedforward calibration and control. The calibration can be performed in a single sweep for each resonator ($\mathcal{O}(N)$) with a specially crafted calibration signal x_{cal} (a). The calibration must be completed in two steps (b and c). Each step completes in $\mathcal{O}(N)$ substeps, if we count that the w vector is measured in a single substep (c). Once the calibration is completed and the crosstalk matrix K is known, the control step uses the stored model to calculate the desired electrical current inputs (d). The overall computational complexity is $\mathcal{O}(N)$. This scheme requires another calibration procedure any time the chip undergoes any environmental variation such as a temperature shift.

Table 3: Efficiency and speed of various index modulation techniques on silicon photonics.


Modulation effect	Speed	Efficiency	Ref.
Volatile mechanisms			
Thermo-optic TiN (a)	$\sim 5.6 \mu\text{s}$	$P_{\pi}L = 6.8 \text{ mW mm}$	[45]
Thermo-optic $\text{N}^+/\text{N}/\text{N}^+$ Si (b)	$\sim \mu\text{s}$	$P_{\pi}L = 0.8 \text{ mW mm}$	[46]
Reverse-biased PN (c)	41 GHz	$V_{\pi}L = 46 \text{ V mm}$	[48]
Graphene SLG (g)	30 GHz ¹	$V_{\pi}L = 28 \text{ V mm}$	[49]
LiNbO ₃ /Si hybrid (f)	70 GHz	$V_{\pi}L = 22 \text{ V mm}$	[50]
III-V MQW/Si hybrid (d)	27 GHz	$V_{\pi}L = 2.4 \text{ V mm}$	[51]
III-V/Si MOS (e)	2.2 GHz	$V_{\pi}L = 0.9 \text{ V mm}$	[52]
ITO MOS (e)	$\sim \text{GHz}^2$	$V_{\pi}L = 0.52 \text{ V mm}$	[53]
Forward-biased PIN (c)	0.5 GHz	$V_{\pi}L = 0.36 \text{ V mm}$	[54]
Non-volatile mechanisms			
PCM Ge ₂ Sb ₂ Te ₅ (h)	0.8 GHz ³	$E_{\pi} = 400 \text{ pJ}$	[10]
PCM Sb ₂ Se ₃ (h)	1.25 MHz	$E_{\pi} = 176 \text{ nJ}$	[55]
Ion implantation	0	0	[56]

Options for phase modulation of silicon waveguides. (a) Thermal tuning with TiN filament; (b) thermal tuning with embedded photoconductive heater; (c) PN/PIN junction across the waveguide for injection and/or depletion modulation; (d) III-V/Si hybrid waveguide; (e) metal-oxide-semiconductor (MOS), where the ‘metal’ is actually an active semiconductor; (f) lithium niobate cladding adds a strong electrooptic effect; (g) 2 single-layer-graphene (SLG) ‘capacitor’; (h) non-volatile phase change material. ¹This bandwidth was not yet shown experimentally. A big challenge is to reduce the contact resistance with Graphene, reducing RC-loading effect. ²Not experimentally shown at high-speed. ³Demonstrated up to 20 MHz. We broke down the mechanisms into volatile and non-volatile based on the reviewer recommendations.

with waveguide-embedded heaters is similar in efficiency [46], but provides a potential feedback signal for weight control [47]. To exploit free-carrier effects, we can directly manipulate carrier concentrations by selectively p- and n-doping the waveguide in a lateral junction [48]. Free-carrier absorption dominates if the junction is forward biased, while free-carrier dispersion dominates if the junction is reverse biased.

Moving beyond pure silicon approaches, one can make hybrid waveguides consisting of a silicon core and additional materials with favorable index modulation

properties that are placed close enough to the core that interact with the evanescent field. Some examples include III-V semiconductors [51], lithium niobate [50], and graphene [49]. These mechanisms are faster and require much less power compared to heaters, but typically provide smaller tuning range before the onset of electrical damage.

Non-volatile mechanisms: Tuning methods based on phase change materials (PCMs) allow weights to retain their values without active power consumption after being set. This is called non-volatile weight (or memory) and provides the most energy efficient weight setting method: note that in Table 3 the efficiency is measured in Joules rather than Watts. This method involves depositing chalcogenide films atop silicon waveguides, e.g., Ge₂Sb₂Te₅ (GST) [10, 57], Ge₂Sb₂Se₄Te₁ (GSST) [58], Sb₂Se₃ [55, 59]. Microring resonator non-volatile weights with low loss have recently been demonstrated with Sb₂Se₃ PCMs [55].

Combining volatile and non-volatile methods: Another important variation source is ambient temperature. Automotive-class devices must have an operating temperature range of -40°C to 125°C . Due to silicon’s large thermo-optic coefficient, this temperature range results in a large resonance drift, albeit no worse than the original fabrication variation (Figure 7). Depending on the final application of the circuit, these two sources of variation can be addressed with different index modulation options from Table 3.

We propose two practical options for addressing variations, including the ones present in automotive temperature ranges. The first is to use post-fabrication trimming to lock resonators to their desired resonance wavelengths at a set operating temperature, which can be chosen, for example, to be above room temperature to avoid condensation. A thermoelectric controller can then be used to stabilize the chip’s resting temperature and fine-tuning can be performed via either microheaters or PN junctions, depending on the required operation speed and actuation range. This greatly simplifies the photonic integrated circuit design, but the use of temperature control adds thermal engineering complexity to the packaging. Another option is to rely on active trimming, in which on-chip microheaters are used to compensate for wide operating temperature ranges. Accompanying this, fine-tuning can be performed via microheaters or PN junctions. This approach requires more control circuitry, but eases the overall burden on packaging [30, 31]. Both approaches can be visualized in Figure 7.

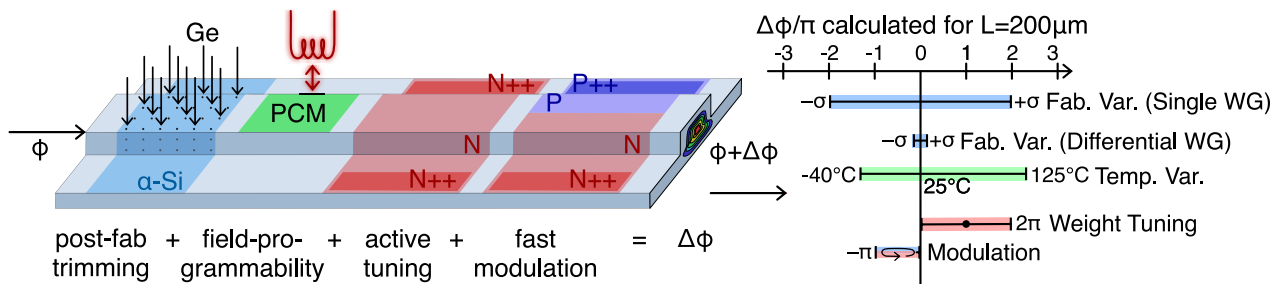


Figure 7: Active trimming and phase modulation strategy. In this example strategy, Ge ion implantation is used as a post-fabrication trimming technique, a PCM is used as a field-programmable non-volatile memory, an N-doped heater is used for weight tuning and configuration, and a standard PN junction is used for fast modulation. The graph on the right shows the relative phase variability from the environment compared to the necessary variation required for modulation and weight configuration. These values were experimentally computed from test structures fabricated through a standard silicon photonics foundry [60].

5.2 Electrical actuation and sensing

As reviewed above, there are multiple physical phenomena that are available for electrically adjusting the refractive index of a waveguide, and consequently the resonance wavelength of an integrated resonator. Since the weight is determined by the distance between a lightwave's wavelength and a resonator's resonance wavelength (Table 2), we would ideally have a sensor that can directly determine resonance wavelength (λ_0). With such a sensor coupled with resonance actuation, we would only need to model the mapping between resonance wavelength and weight (i.e., a Lorentzian function for microring resonators) in order to precisely perform weighting. In reality, it is impractical to measure λ_0 directly because it would require a precision spectrometer.

One practical alternative is to use a temperature sensor [61] surrounding the resonator. This allows absolute temperature stabilization independent from the environment's temperature. Although this provides an efficient mechanism for 'fixing' a weight, it usually requires sourcing very precise currents and reading tiny voltage signals with high dynamic range. So far, it has only been successfully employed to stabilize microring resonators around maximum transmission, which is useful for optical switching but not smoothly-variable weighting.

Another practical alternative is to measure the optical power circulating within a resonator with a photosensitive element, shown in Figure 8. This method only works if the weight value is proportional to the circulating power relative to input power, and if input power is otherwise known. A simple method for measuring circulating power is by lightly doping the resonator's waveguide, which results in a photoconductive resistor. This resistor can be used for both heating and sensing the circulating power,

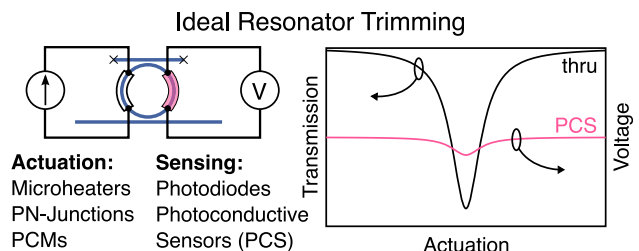


Figure 8: Schematic of a resonator trimming circuit. Effective trimming requires an actuation stage and a sensing stage. Dropped power reaches a maximum when the actuation signal brings the resonator's resonant wavelength to exactly the signal's wavelength.

and was successfully employed in both resonance-locking circuits [46] and weight control [47].

This sensing modality will suffer if the nonheater parasitic resistances¹ are too large (e.g., more than 10 kOhm). Fortunately, when chips are packaged with low-impedance wirebonds and properly designed printed circuit boards (PCBs) all such parasitic resistances become insignificant compared to the heater resistance [62]. We have found that a heater resistance range of 1.5–2.5 kOhm works well. This is because in a typical silicon waveguide, a π phase shift is achieved with the application of ~ 20 mW of heating via the thermo-optic effect. Such resistance therefore helps limit the required voltage to under 10 V and current to under 4 mA, within the ranges of widely available DAC and ADC circuits. Finally, we found that adding noise filtering at the PCB level and shielding cables are key to combat power supply ripple and electromagnetic interference.

¹ These include, e.g., wiring, trace, and electrode contact resistances.

5.3 Feedback control

Equipped with an appropriate sensing element, it is then possible to build a closed-loop control circuit that dynamically responds to external disturbances to a resonator. Similarly, with a sensor for each resonator in a weight bank it becomes possible to correct for disturbances to each resonator individually. This modality of control based on sensed signals is called a feedback (or closed-loop) control scheme. Compared to the feedforward scheme introduced in Section 4, the feedback scheme offers superior accuracy and precision as well as reduced control complexity.

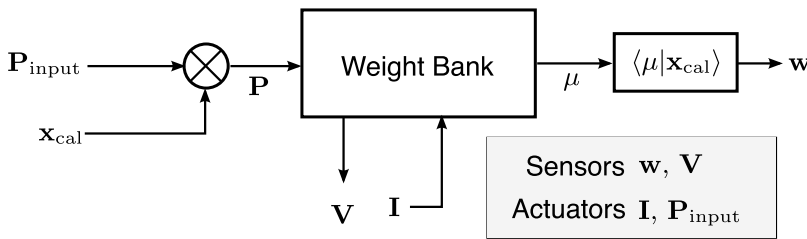
A typical control circuit involves a current-based actuator, voltage-based sensors, and a model that maps the sensing signal to the desired control quantity (Figure 9). Having a sensor for every weight element simplifies the feedback control law from one large model for the weight bank to a collection of independent models for each

resonator. Unlike feedforward control, it is no longer necessary to take into account the thermal crosstalk between resonators or the ambient temperature because the closed-loop circuit corrects for any deviation from the target.

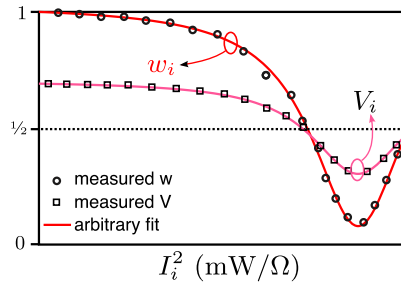
A photoconductive sensor (PCS) provides a voltage signal that is linearly dependent to the transmission function. The transmission function can be translated in wavelength space with an actuator, leading to smoothly-controllable weight values. Rather than modeling how transmission is affected by actuation and other environmental variables, the controller can sense the voltage directly and simply adjust the actuation accordingly.

Despite using a closed-loop/feedback control circuit, each resonator may vary slightly due to fabrication imperfections. In order to compensate for this, all imperfections must be parameterized and calibrated out on a chip-by-chip basis. Fortunately, this model only has to be built once per device. Therefore, it can be done at the manufacturing

(a) Calibration Procedure

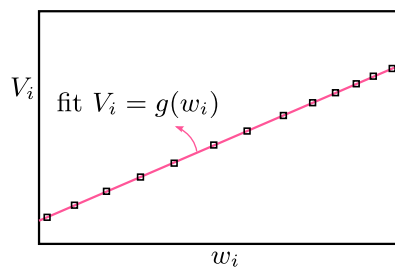


(b) Single Step



For each weight i , set P_i on, sweep I_i and measure w_i and V_i .
Fit curve $V_i = g(w_i)$.

(c)



(d) Control Procedure

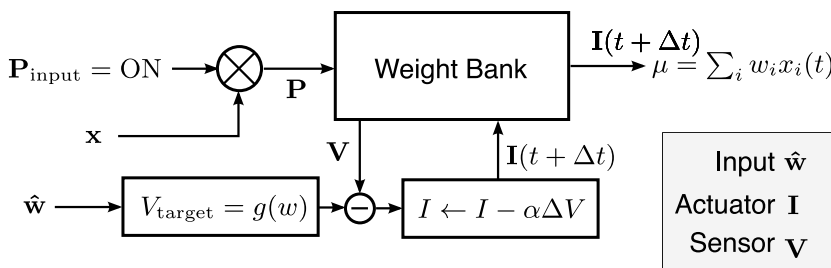


Figure 9: Procedure for feedback calibration and control. The calibration can be performed in a single sweep for each resonator ($\mathcal{O}(N)$) with a specially crafted calibration signal x_{cal} (a). The goal is to fit the relationship between effective weight and the sensor voltage signal (b), which is ideally a linear relationship (c). Once that is established, the control procedure (d) takes a commanded weight, computes the target sensor voltage, and uses a feedback loop to update the actuation electrical current until that voltage is reached. This feedback loop should be resistant to environmental variations so long as the feedback loop can operate at a higher speed than the variations.

facility and stored in a read-only memory co-packaged with the chip.

6 Conclusions

Practical neuromorphic photonic processors employing resonator-based weight banks must be able to achieve high precision and accuracy while being readily manufacturable and deployable in variable/dynamic environments. This is made particularly challenging by the fact that resonators are highly sensitive to both manufacturing variations and environmental disturbances. In order to compensate for this, individual resonator weights must be specifically designed with calibration and control in mind.

In this paper, we have outlined index tuning and sensing mechanisms that can be used for actuation and sensing of weights, as well as feedforward and feedback calibration/control methodologies for use with weight banks. With multiple index tuning mechanisms with varying speed, efficiency, and dynamic range available, system designers can choose which mechanism or combination of mechanisms are suitable based on application needs.

As an example, we outline a fabrication design strategy that can compensate for worst-case fabrication variation ($\Delta\phi = \pm 2\pi$) and employment in automotive-class operating temperature ranges ($-40\text{ }^\circ\text{C}$ – $125\text{ }^\circ\text{C}$). Germanium ion implantation is used for post-fabrication trimming, a phase-change material (PCM) is used to compensate for large temperature variations, an N-doped heater is used for lower-bandwidth weight tuning, and a PN junction is used for fast modulation. Complementing this, photoconductive sensors are used on each resonator in order to enable feedback calibration/control procedures.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: The authors acknowledge Universal Display Corporation and the Office of Naval Research (Grants N00014-18-1-2297 and N00014-20-1-2664) for funding support. S. Bilodeau acknowledges funding from the Fonds de recherche du Québec - Nature et technologies.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix: Microring resonator weight design methodology

In this section, we outline our design methodology for symmetric add–drop microring resonators. We follow the

nomenclature from Ref. [32]. The ring's *add* and *drop* transmission coefficients can be computed analytically using transfer matrices with only four parameters: self-coupling coefficient r (thru port amplitude), round-trip loss coefficient $a \rightarrow ra$ (accounting for the second coupler's coupling loss), cavity length L , and effective index of refraction n_{eff} as a function of wavelength. Each parameter depends on the geometry of the resonator such as waveguide width, the gap between the bus waveguide and the ring, the ring's radius R , and the doping level of the waveguide. r can be computed through finite-difference time domain (FDTD) calculations of the coupler section. a is related to power attenuation coefficient α via $a^2 = \exp(-\alpha L)$, and α is determined by the combination of the loss mechanisms in the ring. This includes the propagation loss from absorption and waveguide imperfections (process-dependent) and bending loss. $L = 2\pi R$ for ring resonators, and the waveguide effective index depends on waveguide material and geometry and is calculated from eigenmode simulations. From the resulting transmission coefficients, the extinction ratios – corresponding to the ratio between maximum and minimum weights – can be computed by evaluating the ratio between the maximum and minimum transmission values on either the *thru* and *drop* ports.

The quantities of interest for weighting can be extracted from the parameters above. The free spectral range is calculated as

$$\text{FSR}(\lambda) = \frac{\lambda^2}{2\pi R} \left(n_{\text{eff}}(\lambda) - \lambda \frac{\partial n_{\text{eff}}}{\partial \lambda} \right)^{-1}.$$

The finesse determines roughly how many WDM channels (or resonators) can share the same bus waveguide

$$F = \pi \arccos \left(\frac{2ar^2}{1 + a^2r^4} \right).$$

The Q -factor approximates the optical rise and fall times of the resonator:

$$Q = \frac{\lambda F}{\text{FSR}}.$$

Finally, the *thru* and *drop* extinction ratios allow calculation of the weighting dynamic range (assuming the index tuning mechanism can be used to tune the resonator on and off-resonance):

$$\text{ER}_{\text{thru}} = \frac{(a+1)^2(1-r^2a)^2}{(a-1)^2(1+r^2a)^2}, \quad \text{ER}_{\text{drop}} = \frac{(1+r^2a)^2}{(1-r^2a)^2}.$$

We used Lumerical FDTD to perform the finite-difference time domain coupler simulations to extract $r(\lambda)$ as a function of geometry, and Lumerical MODE for the curved waveguide eigensolver simulations to obtain

$n_{\text{eff}}(\lambda)$. A Python script implements the above analysis pipeline, and the results can be visualized in a Jupyter Notebook. The Python script and the results of the Lumerical simulations are made accessible online [35].

References

- [1] M. Horowitz, “Computing’s energy problem (and what we can do about it),” in 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, IEEE, 2014, pp. 10–14. [Online]. Available at: <http://ieeexplore.ieee.org/document/6757323/>.
- [2] R. Raina, A. Madhavan, and A. Y. Ng, “Large-scale deep unsupervised learning using graphics processors,” in Proceedings of the 26th Annual International Conference on Machine Learning - ICML ’09, Montreal, Quebec, Canada, ACM Press, 2009, pp. 1–8. [Online]. Available at: <http://portal.acm.org/citation.cfm?doid=1553374.1553486>.
- [3] N. P. Jouppi, C. Young, N. Patil, et al., “In-datacenter performance analysis of a tensor processing unit,” in Proceedings of the 44th Annual International Symposium on Computer Architecture, Toronto ON Canada, ACM, 2017, pp. 1–12. [Online]. Available at: <https://dl.acm.org/doi/10.1145/3079856.3080246>.
- [4] M. Davies, N. Srinivasa, T.-H. Lin, et al., “Loihi: a neuromorphic manycore processor with on-chip learning,” *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [5] F. Akopyan, J. Sawada, A. Cassidy, et al., “TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip,” *IEEE Trans. Comput. Aided Des. Integrated Circ. Syst.*, vol. 34, no. 10, pp. 1537–1557, 2015.
- [6] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, et al., “A million spiking-neuron integrated circuit with a scalable communication network and interface,” *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
- [7] C. Mead, “Neuromorphic electronic systems,” *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.
- [8] P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics*, Boca Raton, CRC Press, 2017.
- [9] A. N. Tait, A. X. Wu, T. Ferreira de Lima, et al., “Microring weight banks,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 22, no. 6, pp. 312–325, 2016.
- [10] C. Ríos, M. Stegmaier, P. Hosseini, et al., “Integrated all-photonic non-volatile multi-level memory,” *Nat. Photonics*, vol. 9, no. 11, pp. 725–732, 2015.
- [11] N. Janosik, Q. Cheng, M. Glick, Y. Huang, and K. Bergman, “High-resolution silicon microring based architecture for optical matrix multiplication,” in 2019 Conference on Lasers and Electro-Optics (CLEO), 2019, pp. 1–2.
- [12] M. Miscuglio and V. J. Sorger, “Photonic tensor cores for machine learning,” *Appl. Phys. Rev.*, vol. 7, no. 3, p. 031404, 2020.
- [13] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, “Holylight: a nanophotonic accelerator for deep learning in data centers,” in 2019 Design, Automation Test in Europe Conference Exhibition (DATE), 2019, pp. 1483–1488.
- [14] F. Sunny, A. Mirza, M. Nikdast, S. Pasricha, and Crosslight, “A cross-layer optimized silicon photonic neural network accelerator,” in 2021 58th ACM/IEEE Design Automation Conference (DAC), 2021, pp. 1069–1074.
- [15] S. Xu, J. Wang, and W. Zou, “Optical convolutional neural network with wdm-based optical patching and microring weighting banks,” *IEEE Photon. Technol. Lett.*, vol. 33, no. 2, pp. 89–92, 2021.
- [16] J. Gu, C. Feng, Z. Zhao, et al., “Squeezelight: towards scalable optical neural networks with multi-operand ring resonators,” in 2021 Design, Automation Test in Europe Conference Exhibition (DATE), 2021, pp. 238–243.
- [17] M. Li and Y. Wang, “An energy-efficient silicon photonic-assisted deep learning accelerator for big data,” *Wireless Commun. Mobile Comput.*, vol. 2020, p. 6661022, 2020.
- [18] Y. Bai, M. Yu, L. Lu, D. Zhang, and L. Zhu, “Quantized photonic neural network modeling method based on microring modulators,” *Opt. Eng.*, vol. 61, no. 6, pp. 1–10, 2022.
- [19] T. Ferreira de Lima, H.-T. T. Peng, A. N. Tait, et al., “Machine learning with neuromorphic photonics,” *J. Lightwave Technol.*, vol. 37, no. 5, pp. 1515–1534, 2019.
- [20] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, et al., “Photonics for artificial intelligence and neuromorphic computing,” *Nat. Photonics*, vol. 15, no. 2, pp. 102–114, 2021.
- [21] C. Huang, S. Fujisawa, T. Ferreira de Lima, et al., “A silicon photonic–electronic neural network for fibre nonlinearity compensation,” *Nat. Electron.*, vol. 4, no. 11, pp. 837–844, 2021.
- [22] G. Frantz, “Digital signal processor trends,” *IEEE Micro*, vol. 20, no. 6, pp. 52–59, 2000.
- [23] U. Ramacher, J. Beichter, W. Raab, et al., “Design of a 1st generation neurocomputer,” in VLSI Design of Neural Networks, U. Ramacher and U. Rückert, Eds., Boston, MA, Springer US, 1991, pp. 271–310. [Online]. Available at: http://link.springer.com/10.1007/978-1-4615-3994-0_14.
- [24] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, “Silicon photonics codesign for deep learning,” *Proc. IEEE*, vol. 108, no. 8, pp. 1261–1282, 2020.
- [25] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, “Cascaded silicon micro-ring modulators for WDM optical interconnection,” *Opt. Express*, vol. 14, no. 20, p. 9431, 2006.
- [26] A. Rahim, T. Spuesens, R. Baets, and W. Bogaerts, “Open-access silicon photonics: current status and emerging initiatives,” *Proc. IEEE*, vol. 106, no. 12, pp. 2313–2330, 2018.
- [27] K. Giewont, S. Hu, B. Peng, et al., “300-mm monolithic silicon photonics foundry technology,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 25, no. 5, pp. 1–11, 2019.
- [28] L. Chrostowski, H. Shoman, M. Hammond, et al., “Silicon photonic circuit design using rapid prototyping foundry process design kits,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 25, no. 5, pp. 1–26, 2019.
- [29] K. Padmaraju and K. Bergman, “Resolving the thermal challenges for silicon microring resonator devices,” *Nanophotonics*, vol. 3, nos 4-5, pp. 269–281, 2014.
- [30] C. T. DeRose, M. R. Watts, D. C. Trotter, D. L. Luck, G. N. Nielson, and R. W. Young, “Silicon microring modulator with integrated heater and temperature sensor for thermal

- control,” in Conference on Lasers and *Electro-Optics* 2010, San Jose, California, OSA, 2010, p. CThJ3. [Online]. Available at: <https://www.osapublishing.org/abstract.cfm?URI=CLEO-2010-CThJ3>.
- [31] S. Grillanda, M. Carminati, F. Morichetti, et al., “Non-invasive monitoring and control in silicon photonics using CMOS integrated electronics,” *Optica*, vol. 1, no. 3, p. 129, 2014.
- [32] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, et al., “Silicon microring resonators,” *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, 2012.
- [33] N. Eid, R. Boeck, H. Jayatilleka, L. Chrostowski, W. Shi, and N. A. Jaeger, “Fsr-free silicon-on-insulator microring resonator based filter with bent contra-directional couplers,” *Opt. Express*, vol. 24, no. 25, pp. 29 009–29 021, 2016.
- [34] H. Zhou, C. Qiu, X. Jiang, et al., “Compact, submilliwatt, 2×2 silicon thermo-optic switch based on photonic crystal nanobeam cavities,” *Photon. Res.*, vol. 5, no. 2, p. 108, 2017.
- [35] S. Bilodeau, T. Ferreira de Lima, “Microring Resonator Weight Design” [Jupyter Notebook], 2022. <https://doi.org/10.5281/zenodo.6419611>. Available at: <https://github.com/lightwave-lab/microring-resonator-weight-design/releases/tag/nanoph-2022>.
- [36] M. Nedeljkovic, R. Soref, and G. Z. Mashanovich, “Free-carrier electrorefraction and electroabsorption modulation predictions for silicon over the 1–14 μm infrared wavelength range,” *IEEE Photonics J.*, vol. 3, no. 6, pp. 1171–1180, 2011.
- [37] A. N. Tait, T. Ferreira de Lima, M. P. Chang, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Application regime and distortion metric for multivariate RF photonics,” in 2017 IEEE Optical Interconnects Conference (OI), 2017, pp. 25–26. Available at: <https://doi.org/10.1109/oic.2017.7965513>.
- [38] C. Demirkiran, F. Eris, G. Wang, et al., “An electro-photonics system for accelerating deep neural networks,” in *arXiv:2109.01126 [cs]*, 2021, [Online]. Available at: <http://arxiv.org/abs/2109.01126>.
- [39] N. C. Harris, J. Carolan, D. Bunandar, et al., “Linear programmable nanophotonic processors,” *Optica*, vol. 5, no. 12, pp. 1623–1631, 2018.
- [40] K. Williams, R. Esman, and M. Dagenais, “Effects of high space-charge fields on the response of microwave photodetectors,” *IEEE Photon. Technol. Lett.*, vol. 6, no. 5, pp. 639–641, 1994.
- [41] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Multi-channel control for microring weight banks,” *Opt. Express*, vol. 24, no. 8, p. 8895, 2016.
- [42] S. K. Selvaraja, W. Bogaerts, P. Dumon, D. Van Thourhout, and R. Baets, “Subnanometer linewidth uniformity in silicon nanophotonic waveguide devices using CMOS fabrication technology,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 16, no. 1, pp. 316–324, 2010.
- [43] L. Chrostowski, X. Wang, J. Flueckiger, Y. Wu, Y. Wang, and S. T. Fard, “Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits,” in *Optical Fiber Communication Conference*, San Francisco, California, OSA, 2014, p. Th2A.37. [Online]. Available at: <https://www.osapublishing.org/abstract.cfm?URI=OFC-2014-Th2A.37>.
- [44] Z. Lu, J. Jhoja, J. Klein, et al., “Performance prediction for silicon photonics integrated circuits with layout-dependent correlated manufacturing variability,” *Opt. Express*, vol. 25, no. 9, pp. 9712–9733, 2017.
- [45] M. Jacques, A. Samani, E. El-Fiky, D. Patel, Z. Xing, and D. V. Plant, “Optimization of thermo-optic phase-shifter design and mitigation of thermal crosstalk on the SOI platform,” *Opt. Express*, vol. 27, no. 8, p. 10456, 2019.
- [46] H. Jayatilleka, K. Murray, M. Á. Guillén-Torres, et al., “Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters,” *Opt. Express*, vol. 23, no. 19, p. 25084, 2015.
- [47] A. N. Tait, H. Jayatilleka, T. Ferreira De Lima, et al., “Feedback control for microring weight banks,” *Opt. Express*, vol. 26, no. 20, pp. 26 422–26 443, 2018.
- [48] D. Patel, S. Ghosh, M. Chagnon, et al., “Design, analysis, and transmission system performance of a 41 GHz silicon photonic modulator,” *Opt. Express*, vol. 23, no. 11, p. 14263, 2015.
- [49] V. Soriano, M. Midrio, G. Contestabile, et al., “Graphene-silicon phase modulators with gigahertz bandwidth,” *Nat. Photonics*, vol. 12, no. 1, pp. 40–44, 2018.
- [50] M. He, M. Xu, Y. Ren, et al., “High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbit/s and beyond,” *Nat. Photonics*, vol. 13, no. 5, pp. 359–364, 2019.
- [51] H.-W. Chen, J. D. Peters, and J. E. Bowers, “Forty Gb/s hybrid silicon Mach-Zehnder modulator with low chirp,” *Opt. Express*, vol. 19, no. 2, p. 1455, 2011.
- [52] T. Hiraki, T. Aihara, K. Hasebe, et al., “Heterogeneously integrated III-V/Si MOS capacitor Mach-Zehnder modulator,” *Nat. Photonics*, vol. 11, no. 8, pp. 482–485, 2017.
- [53] R. Amin, R. Maiti, C. Carfano, et al., “0.52 V-mm ITO-based Mach-Zehnder modulator in silicon photonics,” *APL Photonics*, vol. 3, no. 12, p. 126104, 2018.
- [54] W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, “Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator,” *Opt. Express*, vol. 15, no. 25, p. 17106, 2007.
- [55] C. Ríos, Q. Du, Y. Zhang, et al., “Ultra-compact nonvolatile phase shifter based on electrically reprogrammable transparent phase change materials,” in *arXiv:2105.06010 [cond-mat, physics:physics]*, 2022, arXiv: 2105.06010. [Online]. Available at: <http://arxiv.org/abs/2105.06010>.
- [56] M. M. Milosevic, X. Chen, W. Cao, et al., “Ion implantation in silicon for trimming the operating wavelength of ring resonators,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 24, no. 4, pp. 1–7, 2018.
- [57] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, “All-optical spiking neurosynaptic networks with self-learning capabilities,” *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.
- [58] Y. Zhang, J. B. Chou, J. Li, et al., “Broadband transparent optical phase change materials for high-performance nonvolatile photonics,” *Nat. Commun.*, vol. 10, no. 1, p. 4279, 2019.
- [59] M. Delaney, I. Zeimpekis, D. Lawson, D. W. Hewak, and O. L. Muskens, “A new family of ultralow loss reversible phase-change materials for photonic integrated circuits: Sb_2S_3 and Sb_2Se_3 ,” *Adv. Funct. Mater.*, vol. 30, no. 36, p. 2002447, 2020.

- [60] L. Chrostowski, X. Wang, J. Flueckiger, Y. Wu, Y. Wang, and S. T. Fard, "Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits," in *Optical Fiber Communication Conference*, Washington, D.C., OSA, 2014, p. Th2A.37. [Online]. Available at: <https://www.osapublishing.org/abstract.cfm?URI=OFC-2014-Th2A.37>.
- [61] S. Saeedi and A. Emami, "Silicon-photonic PTAT temperature sensor for micro-ring resonator thermal stabilization," *Opt. Express*, vol. 23, no. 17, p. 21875, 2015.
- [62] C. Huang, S. Bilodeau, T. Ferreira de Lima, et al., "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photonics*, vol. 5, no. 4, p. 040803, 2020.