Quantifying Power in Silicon Photonic Neural Networks

Alexander N. Tait[*,†]

Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada

(Received 10 August 2021; revised 27 November 2021; accepted 15 February 2022; published 18 May 2022)

Due to challenging efficiency limits facing conventional and unconventional electronic architectures, information processors based on photonics have attracted renewed interest. Research communities have yet to settle on definitive techniques to describe the performance of this class of information processors. Photonic systems are different from electronic ones, and the existing concepts of computer performance measurement cannot necessarily apply. In this paper, we quantify the power use of photonic neural networks with state-of-the-art and future hardware. We derive scaling laws, physical limits, and platform platform performance metrics. We find that overall performance takes on different dominant scaling laws depending on scale, bandwidth, and resolution, which means that energy efficiency characteristics of a photonic processor can be completely described by no less than seven performance metrics over the range of relevant operating domains. The introduction of these analytical strategies provides a much needed foundation and reference for quantitative roadmapping and commercial value assignment for silicon photonic neural networks.

DOI: 10.1103/PhysRevApplied.17.054029

I. INTRODUCTION

The computational requirements and energy expenditures of machine learning are so staggering that its prevalence is becoming a climate issue [1]. In the pursuit of energy efficiency, massively distributed hardware has been developed [2]. These pursuits have come to include nondigital signaling [3,4] and post-CMOS platforms [5], including photonic platforms [6,7]. One of the value propositions of photonic neural networks and vector-matrix multipliers (VMMs) is reduced energy use in performing linear operations.

It is essential to have a rigorous understanding of power scaling laws and limits in order to support that proposition. This understanding is foundational to system design and roadmapping, recognizing what the limiting factors are, and prioritizing technological developments that address those factors. In this paper, we study the energy consumption and computational efficiency of two silicon photonic neural-network architectures [6,8] described in Fig. 1. We attempt to answer the key questions of how efficient they are in the worst case, how efficient they could be, and which technologies have the greatest impact on getting there.

Energy efficiency of photonic neural networks has been studied in prior works in depth [9,10]. Other works have proposed linear scaling laws that are only applicable to small-scale operating points [6,11–13], neglecting quadratic scaling laws. Some of these scaling laws have been extrapolated with optimism to make predictions that idealized optical systems can perform multiply accumulate operations (MACs) for free in the limit of large matrices. We refute these predictions using energy-conservation arguments. Large matrices are subject to scaling laws that dictate that MAC efficiency approaches finite values. This finding is true for both architectures analyzed: multiwavelength [based on wavelength-division multiplexed (WDM) weight banks] and coherent (based on Mach-Zehnder interferometer meshes). These two architectures, despite distinct theories of operation (Fig. 1), are found to share several identical scaling laws.

Despite what might be seen as pessimistic performance predictions (relative to past work), we find that photonic neural networks and VMMs can be highly competitive compared to state-of-the-art electronics. It is worth noting that improving energy efficiency is not the only value proposition of photonic neural networks. Their bandwidth and latency can enable new real-time applications that are unaddressable by foreseeable electronic processors and are, therefore, compelling regardless of energy efficiency. These alternative application-based opportunities are left out of scope, which is here restricted to power use and its physics-based quantification.

This paper takes a strategy of identifying invariant quantities that grant insight into the interplay of the various power contributors (i.e., weight tuning power, laser pump power, etc.). Total power use (P) can be described as a

[*]alex.tait@queensu.ca

[†]Previously at Physical Measurement Laboratory, National Institute of Standards and Technology, Boulder, CO 80305, USA.

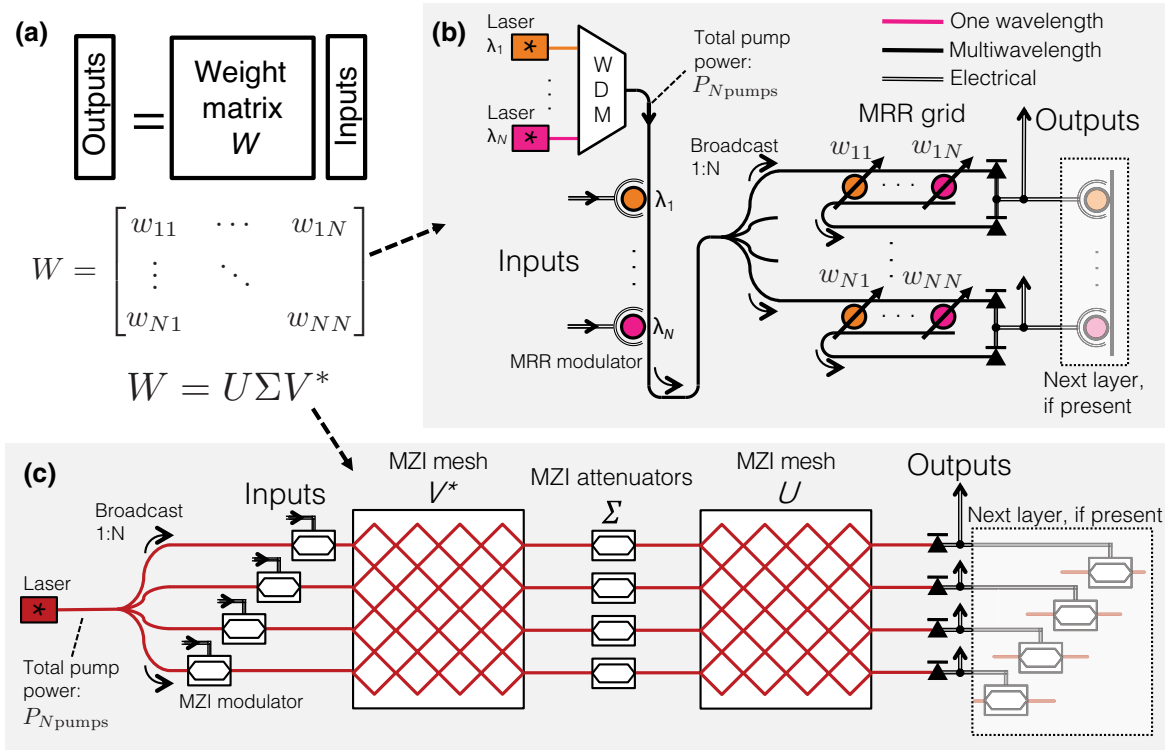

FIG. 1. Silicon photonic neural-network architectures. Optical pumps and electrical inputs and outputs are shown. (a) The vector-matrix operation that is central to neural interconnects. The weight matrix, W, can be broken down into elements, w_{ij}, or into two unitary (U, V^*) and one diagonal (Σ) matrices. (b) A multiwavelength broadcast-and-weight network [8]. Laser pumps are wavelength-division multiplexed. Each wavelength ($\lambda_1 \ldots \lambda_N$) is modulated by one electrical input, and all are broadcast. Microring weights in a grid layout each represent one element of the weight matrix. Balanced photodiodes detect the sum of each row. MRR color corresponds to the wavelength it acts upon. (c) A silicon coherent nanophotonic network [6]. One laser pump provides power to every input beam. Each beam is modulated by one input. Tunable MZI meshes implement the unitary transforms. The singular, diagonal matrix corresponds to an elementwise multiplication, which is implemented by an array of attenuators. The optical output is converted back to the electrical domain. Depending on use case, the output can cascade to another layer (feedforward network, shown in dashed boxes), connect back to the original inputs (recurrent network), or merely be digitized and used elsewhere (vector-matrix multiplier). To function, both require some amount of optical power: $P_{N\text{pumps}}$. Low-bandwidth electronics for weight configuration are not shown. MRR, microring modulator; MZI, Mach-Zehnder interferometer.

sum of polynomials of the form $P \propto AN^x f^y z^B$, where each polynomial term corresponds to one power contributor. N is the number of channels or neurons, f is bandwidth, and B is resolution in effective bits. Each contributor is described by the scaling polynomials (x, y, z) and the invariant coefficient (A), also referred to as a metric. The power contributors dominate in different regimes of (N, f, B). By deconstructing these regimes, this approach provides more detailed insight than approaches that simply calculate the overall power or approaches that do not account for all of the contributors.

The aspects of this paper can be organized by section. Section II introduces a quantification of expected power to counteract fabrication variation in a square matrix of microring resonator (MRR) weights. We propose a path to reduce this power by 4 orders of magnitude by pairing two technologies. Section III derives resolution-determined scaling laws and closed-form expressions of their energetic scaling coefficients. Prior works have studied physical sources of noise in a single neuron [14] and network scaling behavior in an abstract sense [15]. In multichannel photonic information hardware [16], in general, there is a pronounced need for physics-based derivation of resolution concepts. Previously unrealized insights about noise in photonic information processing result, including a shot noise limit on MAC efficiency and a hard limit on bandwidth.

Section IV analyzes gain-determined power as set by cascadability and/or digitization requirements. Arguments based only on energy conservation and functional generality result in a scaling law that refutes certain optimistic scaling laws conceived of in some prior analyses. Section V examines $O/E/O$ transduction in analog neurons and proposes a role for photoelectric amplifiers. Scaling laws and energetic coefficients relating to noise, gain, and detection are found to be nearly identical between

multiwavelength MRR-based [Fig. 1(b)] and coherent MZI-based [Fig. 1(c)] architectures for both photonic neural networks and VMMs. These similarities are surprising because of the fundamental differences in how the architectures employ different properties of light. For a glossary of variables used in these sections, see the Supplemental Material [17]. Finally, Sec. VI compares all of these contributors in terms of dominant regimes, in the process providing a roadmap for device technologies and a walkthrough of thought processes for future system design.

II. WEIGHT CONTROL

Photonic weighted addition schemes depend on the ability to configure the transmission state of passive elements. So far, many proposals for actively programming weights employ thermooptic tuning to achieve the necessary phase shifts. The required power breaks down into a static and a configurable component and is proportional to the number of weights.

$$P_{\text{wei}} = N^2 \left(P_{\text{lock}} + P_{\text{conf}}\right). \tag{1}$$

The static component is needed to lock MRR weights onto their resonances, counteracting fabrication variations. The configurable component is the power needed to tune the MRR on and off resonance in order to program the desired weight value.

A. Weight locking power

Weight locking power is the electrical power needed to bias a weight. MZI architectures do not need biasing because MZIs are wavelength independent. MRRs, on the other hand, must be held close to the on-resonance condition with a WDM carrier. Fabrication nonidealities result in a wide variability in as-fabricated resonant wavelength. The only tuning effects strong enough to counteract this variability are thermal. Thermal locking dissipates a large and static amount of heat on chip. In Ref. [18], it was found that, in some operating regimes of a photonic network-on-chip, static MRR heating accounts for up to 80% of total power. Static locking can dominate in simple communication links, such as in Ref. [19] (80% of total), but not always [20] (0.01%–23%, depending on temperature).

On a silicon-on-insulator wafer, silicon thickness can vary by up to 5% [21], and this variation has a strong effect on as-fabricated MRR resonant wavelengths. At most one free spectral range (FSR) of tuning range is needed to lock one resonance onto an absolute wavelength target; however not all MRRs in the array must lock to an absolute wavelength. Instead, they are locked to a specified wavelength offset from the neighboring MRR. Nearby locations on the wafer are likely to have a similar silicon thickness. This means that the standard deviation of resonance offset is correlated with the distance between MRRs. For resonators spaced by r [mm], standard deviation of wavelength offset is

$$\sigma_{[\text{FSR}]}(r) = \sigma^{(0)}_{[\text{FSR}]} + \sigma^{(1)}_{[\text{FSR}]} r, \tag{2}$$

where the subscript [FSR] means in free spectral range units. Reference [22] measured these parameters for the IME A*STAR process to be $\sigma^{(0)}_{[\text{FSR}]} = 0.050$ and $\sigma^{(1)}_{[\text{FSR}]} = 0.060\ \text{mm}^{-1}$ with the FSR at 7 nm in that work. Variances can also be stated in wavelength units (denoted with subscript [λ]) by multiplying by the FSR in wavelength units.

We introduce a term Ω to indicate the expected value of resonant shift per MRR needed to bring a square array of MRRs onto resonance.

$$\Omega(N) = \min\left[\sigma_{[\text{FSR}]}(Nd), \frac{1}{2}\right], \tag{3}$$

$$P_{\text{lock}} = K\Omega(N), \tag{4}$$

where d is the MRR pitch, and Nd is the side length of the square MRR array. Ω is in FSR units. K is the tuning efficiency in mW per FSR units. The MRR pitch, d, is taken here to be 20 μm.

B. Weight configuration power

Weight configuration power is used to program the weight value. Applying heat tunes the MRR from on resonance (weight −1) to slightly off resonance (weight +1). Supposing that tuning over a FWHM is required, this power is

$$P_{\text{conf}} = \frac{K}{2\mathcal{F}}, \tag{5}$$

where $\mathcal{F}$ is finesse. We state K in FSR units, so finesse converts it to FWHM units. The factor of 2 results from averaging over the range of possible states from on resonance to off resonance by one FWHM. We can approximate finesse as roughly 100 for typical silicon MRRs, although optimized traveling wave resonators have achieved finesse up to 1140 [23]. The vertical junction depletion modulators discussed below had a finesse of 277. Typical values for K are given in Table I.

The MRR resonance has a sharp wavelength dependence, meaning that—once locked—there is a small incremental power needed to configure the weight. If we are unable to control where the resonance falls as fabricated, then locking power will be greater than configuration power by a factor of the finesse.

1. MZI weight configuration

The opposite power balance is found in MZI mesh architectures. MZIs are less sensitive to fabrication variation,

TABLE I. Platform values for weight configuration.

Name	Variable	Value	Description
Variation	$\sigma^{(0)}_{[\mathrm{FSR}]}$	0.050	Resonance offset SD, minimum (FSR units) [22]
		0.0055	Reduced MRR variability using trimming [28]
Covariation	$\sigma^{(1)}_{[\mathrm{FSR}]}$	0.060 mm^{-1}	Resonance offset SD, distance dependence (FSR units) [22]
		0 mm^{-1}	Reduced MRR variability using trimming [28]
MRR tuning efficiency	K	28 mW/FSR	Embedded N-doped heater [39]
		2.4 mW/FSR	Trench etched [25,26]
		0.13 mW/FSR	Vertical junction depletion [20]
MZI tuning efficiency	P_π	10 mW/π	Baseline thermal phase shifter [40]
		1.2 mW/π	Trench etched [25]
		100 nW/π	Barium titanate [30]
		< 100 nW/π	MEMS phase shifter [37,38]

and they are correspondingly less sensitive to desirable tuning effects. MZI tuning power is quantified by π power, P_π: the power needed for a thermal phase shifter to impart an optical phase shift of π. In the above architecture from Ref. [6], there are four phase shifters per matrix element whose average expected power is halfway between minimum phase (0) and maximum phase (π) states.

$$P_{\mathrm{conf,MZI}} = 2P_{\pi,\mathrm{MZI}}. \tag{6}$$

MZI thermal π powers are on the order of of 10 mW [24]. Prior work on MZI meshes has calculated system power to be approximately $1\ \mathrm{mW} \cdot N$ [6] or approximately $100\,\mathrm{mW} \cdot N$ [11], but then they neglected weight configuration power, which would severely dominate at $10\,\mathrm{mW} \cdot N^2$.

MZI meshes do not need static weight locking power; however, MZI configuration power and MRR locking power are both in the mW range. They stem from different needs. The first is due to an essential need to program the weights, and the second is due to fabrication nonideality. This means that MZI meshes have a fundamental need for strong tuning effects, while MRR weights can reduce tuning power by addressing fabricated resonance variability.

C. Weight reconfiguration energy

Weight reconfiguration energy is additional energy needed to change weights on a fast timescale, which is distinct from weight tuning. Typically, fast tuning requires nonthermal tuning, such as depletion modulators that do not draw continuous power. Reconfiguration energy is

$$P_{\mathrm{reconfig}} = N^2 f_{\mathrm{reconfig}} \times E_{\mathrm{reconfig}}, \tag{7}$$

where E_{reconfig} is the same as the energy-per-bit value when considering each tuning element as a modulator. The reconfiguration rate, f_{reconfig}, is an expected, averaged rate that is less than the maximum reconfiguration bandwidth. f_{reconfig} is highly application dependent.

In many neural networks, the reconfiguration of weights happens at much slower timescales than signal timescales. In those cases, the power needed to change weights can generally be neglected. In other applications, such as general VMMs, weights must change at timescales similar to the signals. Since it is application dependent, we largely leave reconfiguration energy from the rest of the analysis.

D. Foreseeable technology

Resonator locking metrics are listed in Table I. More efficient thermal tuning improves both MRRs and MZIs via K and P_π [25,26]. Resonator variability reduction techniques—either prefabrication [27] or postfabrication [28,29]—improves MRRs via Ω. More significant reductions to tuning power could be realized by moving beyond thermal tuning effects.

A key consideration for nonthermal tuning effects is their range, as opposed to efficiency, specifically whether they provide a complete tuning range: $\geq \pi$ phase shift for MZIs, or $\geq$ FSR for MRRs. Tuning elements based on Pockels effect in barium titanate (BTO) can provide a full range [30], although BTO tuners are typically longer than thermal tuners, increasing propagation loss per neuron. Recent work has shown that a 220-μm shifter is feasible [31,32]. Another promising approach is offered by nonvolatile tuners based on phase-change materials, which can eliminate static locking and configuration powers entirely [33,34]. Strong, low-power tuning can also be achieved with photonic MEMS [35], waveguide structures suspended in air that deform in an applied electrostatic field. Photonic MEMS for MZI phase tuning are low loss and faster than thermal effects (100 kHz–10 MHz) [35]. Mechanical tuning techniques have been explored for both MRRs [36] and MZIs [37,38].

Depletion tuning in a p-n junction waveguide is too weak of an effect to achieve a full tuning range; however, for MRRs, variability reduction may be combined with depletion tuning. If the variation reduction technique is accurate enough, it crosses a threshold at which

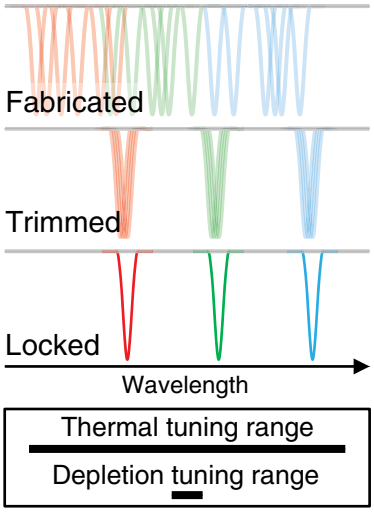

FIG. 2. Concept of using trimming to reduce resonator variation below the depletion-mode tuning range. Color means the channel to which a particular MRR will be assigned. As-fabricated resonances fall over some range that is less than the thermal tuning range but greater than the depletion tuning range. Trimming reduces variation, such that depletion tuning is sufficient to lock them to consistent wavelengths.

point depletion-based locking becomes viable, a concept introduced in Fig. 2. This technique could reduce locking power by 5 orders of magnitude. It may be favorable because depletion tuning devices are highly developed and already present on mainstream silicon photonic platforms. To illustrate the calculation of the MRR weak-tuning threshold, we can consider variability values obtained by Alipour *et al.* [28] and tuning-range values for a vertical junction microdisk obtained by Timurdogan *et al.* [20]. The devices had similar geometries, fundamental modes, and FSRs, making them easier to compare and potentially compatible. In Ref. [20], a 1.1-V bias resulted in a 270-pm resonance shift and 0.7-μA leakage current. Given the radius of 2.4 μm, this leads to an extrapolated FSR efficiency of $K = 0.13$ mW/FSR. Efficiency degraded precipitously above 270 pm due to reverse leakage current, even though the device survived a 680-pm resonance shift. In Ref. [28], the microtoroids also had a FSR of 45 nm. Their initial resonance std. dev. (SD) of 290 pm was reduced to 25 pm after postfabrication trimming, well within the efficient tuning range in [20]. A direction for device research will be demonstrating variability reduction together with depletion modulation in resonators.

III. SIGNAL RESOLUTION

In this section, we consider the laser pump power needed to achieve a certain signal frequency, f , and resolution, B, in effective number of bits. The signal frequency means the bandwidth of the waveform modulating optical power envelopes, which encodes the values of analog variables. We extend upon analog photonic link theory from Ref. [41] to derive analytical power use expressions and extend it to multiple channels. The analysis—besides relative intensity noise—applies identically to multiwavelength and coherent architectures because they are both based on power-encoded signals.

A. Analog photonic links

Further analysis rests on an understanding of laser power, resolution, gain, and nonlinearity in a single analog photonic link (APL) consisting of a modulator connected to a detector. Every photonic processor, including $N \times N$ weight matrices, has an electrical input to a modulator, a linear subsystem, and an electrical output from a photodetector. This analysis was performed by Marpaung in Ref. [41]. Since it is foundational, basic APL theory is rederived in Appendix A.

A key feature of APLs is the existence of operating regimes where different sources of noise dominate the signal-to-noise ratio, or, more precisely, the spurious-free dynamic range (SFDR). In the low power regime, thermal noise originating in the photodetector dominates because the received signal is weak. At the highest powers, relative intensity noise (RIN) from the laser dominates. These regimes and the net SFDR are plotted in Fig. 8, derived in the Appendix. Whereas Marpaung sought to maximize SFDR performance in a single-channel link, here, we are interested in energy efficiency of a multichannel link. Subsequently, we extend APL theory to multichannel architectures.

In digital systems, each added bit increases resolution in proportion, so the required energies are a polynomial function of bit resolution. On the other hand, the power needed to generate an analog signal scales exponentially with its bit resolution [42]. In other words, the endeavor for energy efficiency with nonlinear E/O transfer functions rapidly becomes futile around 6 bits and then strictly nonviable around 10 bits. Striking this balance, we focus on the 2–6 bit regime of APLs. For reference, the TrueNorth neuromorphic electronic processor uses 4-bit weights [43].

B. Single-channel regimes

Appendix A rederives SFDR as a function of optical pump power subject to three sources of noise: thermal, shot, and relative intensity. Each noise component is critical to consider because they scale differently and therefore dominate in different operating regimes. Here, we extend this analysis to arrive at coefficients relating pump power to frequency and effective bits. Typical values of these coefficients are found in Table II.

Unlike for digital signals, analog resolution is not a number of wires or serial bit slots; however, analog signal resolution can be stated in terms of an effective number of

TABLE II. Laser pump-power metrics for single-channel analog photonic links.

Regime	Coefficient[a]	$B = 2$ bit	$B = 4$ bit	$B = 6$ bit	$B = 8$ bit	Units
Thermal	$J^*(B, R_b)\|_{R_b=50\ \Omega}$	250×10^{-3}	2.0	16	130	$\mathrm{nW}/\sqrt{\mathrm{Hz}}$
Thermal	E_{thrm}	820×10^{-3}	6.5	52	420	fJ
Shot	E_{shot} (typical)	24×10^{-3}	1.5	96	6.2×10^3	fJ
Shot	E_{shot} (physical limit)	15×10^{-3}	0.96	61	3.9×10^3	fJ
RIN	F_{RIN}	110×10^3	1.7×10^3	26	0.41	GHz

[a] $R_{\mathrm{PD}} = 0.8$ A/W, $C_{\mathrm{pd}} = 35$ fF, $M = 1$, $T = 300$ K, $\lambda = 1550$ nm.

bits corresponding to an equivalent digital signal. Analog effective number of bits is related to SFDR, which, roughly speaking, is the ratio of maximum to minimum resolvable signals.

The conversion of SFDR spectral density to bits is

$$B[\mathrm{bits}] = \frac{1}{10\log 2}\frac{\mathrm{SFDR[dB]} - 10\log(3/2)}{2} \tag{8}$$

$$= \frac{\mathrm{SFDR} - 1.76}{6.02}, \tag{9}$$

where B is resolution in effective number of bits, and SFDR is defined in more detail in the Appendix. The first term converts base 10 to base 2, and the factor of 2 comes from the fact that SFDR is an electrical power and resolution is measured in terms of voltage. The 1.76 arises due to fundamental quantization error.

SFDR is an appropriate definition of effective resolution because it accounts for nonlinear distortion in optical modulators in addition to noise corruptions. There are cases where nonlinear distortion can be compensated, as considered in Sec. VII D, although compensation requires sophisticated additional hardware. The SFDR from system parameters is derived in the Appendix. Here, we convert SFDR to effective bits and connect it to link power.

1. Thermal regime

Thermal noise (also known as Johnson-Nyquist noise) is caused by the random motions of electrons in the receiver circuitry. From Eq. (A20), the SFDR in the thermal regime is

$$\mathrm{SFDR}\ [\mathrm{dB\,Hz^{2/3}}] = \frac{2}{3}\left[20\log P_{\mathrm{1pump}} + 10\log\left(\frac{R_b}{k_B T}\frac{\eta_{\mathrm{net}}^2 M^2 R_{\mathrm{PD}}^2}{4}\right)\right], \tag{10}$$

where P_{1pump} is the optical power emitted by one pump laser; R_b is receiver impedance; k_B is the Boltzmann constant; T is temperature; η_{net} is passive insertion loss; M is avalanche gain, if present; and R_{PD} is detector responsivity. This equation represents a ratio of signal to noise per unit of spectrum. Combining this thermal SFDR equation with the effective bits equation, Eq. (9), results in an expression of pump power needed for a given bit value

$$10\log P_{\mathrm{1pump}} = \frac{30}{2}\log 2 \times B + \frac{30}{4}\log(3/2) + \frac{10\log f}{2} - \frac{10}{2}\log\left(\frac{R_b}{k_B T}\frac{\eta_{\mathrm{net}}^2 M^2 R_{\mathrm{PD}}^2}{4}\right). \tag{11}$$

In linear units, this one-channel power expression is

$$P_{\mathrm{1pump}}(B, f) = \sqrt{f} \times \frac{J^*(B, R_b)}{\eta_{\mathrm{net}}}, \tag{12}$$

$$\text{where } J^*(B, R_b) \equiv 2^{(3/2)B}\left(\frac{3}{2}\right)^{3/4}\sqrt{\frac{4k_B T}{R_b}}\frac{1}{MR_{\mathrm{PD}}}, \tag{13}$$

where we introduce a term, $J^*(B, R_b)$, that links power, frequency, and resolution in the thermal noise regime for a particular receiver impedance. The link loss, η_{net}, is separated because it will later become a function of network size. J^* has units of energy-per-root-frequency.

The impedance, R_b, is an argument because it is a free design parameter. It can be designed to take on a wide range of resistance values, although its value is fixed at fabrication for a particular chip. When a network is meant to operate at a particular bandwidth, f, there is an optimal design of the junction impedance such that it allows no more than the required signal bandwidth: $R_b = (2\pi f C_{\mathrm{pd}})^{-1}$. The capacitance is not a free design parameter, rather a circuit parasitic determined by the layer thicknesses and device sizes available on a particular fabrication platform. This means there is an additional relation for the optimal design:

$$E_{\mathrm{thrm}}(B) \equiv J^*(B, R_b)\big|_{R_b=(2\pi f C_{\mathrm{pd}})^{-1}} \tag{14}$$

$$= 2^{(3/2)B}\left(\frac{3}{2}\right)^{3/4}\sqrt{8\pi k_B T}\frac{\sqrt{C_{\mathrm{pd}}}}{MR_{\mathrm{PD}}}, \tag{15}$$

where we introduce a term, $E_{\mathrm{thrm}}(B)$, describing the laser pump power needed to support an APL of a given frequency and resolution, supposing an optimal receiver design. Like J^*, this term links power, frequency, and resolution in the thermal regime; unlike, J^*, it has units of

energy—hence our choice of the variable E—resulting in an intuitive power relation:

$$P_{1\text{pump}}(B,f) = f\,\frac{E_{\text{thrm}}(B)}{\eta_{\text{net}}}, \tag{16}$$

where $P_{1\text{pump}}$ is the pump laser power for a single APL, and η_{net} is the transmission efficiency of the APL.

This equation has several notable features that will carry through to the multichannel system. Firstly, power scales exponentially with the number of bits, which is characteristic of analog signaling. Strikingly, the power-resolution scaling rate of $15\log 2 = 4.5$ dB/bit is less than that of any analog electrical link: $20\log 2 = 6.0$ dB/bit. This difference is explained by the fact that output electrical signal power is the square of the received photocurrent and thus optical pump power. The quadratic relation between signal power and supply power also explains the square-root dependence on bandwidth in Eq. (12). For a fixed receiver resistance [Eq. (12)], the photonic system transmits more information per Joule as its bandwidth increases; however, for a resistance that varies optimally with operating bandwidth [Eq. (16)], the information per Joule does not vary. Finally, the avalanche photodiode (APD) gain, M, plays a prominent role. We discuss APDs as a key technology below.

2. Shot-noise regime

Shot noise is due to the randomness in the detection times of quantized photons. From Eq. (A17), the SFDR in the shot-noise regime is

$$\text{SFDR}\ [\text{dB Hz}^{2/3}] = \frac{2}{3}\left[10\log P_{1\text{pump}} + 10\log\left(\frac{\eta_{\text{net}}R_{\text{PD}}}{qF_A}\right)\right], \tag{17}$$

where q is electron charge, and F_A is excess noise introduced by an APD, if present. Combining again with Eq. (9), we arrive at power needed for a given bit value,

$$10\log P_{1\text{pump}} = 30\log 2 \times B + \frac{30}{2}\log(3/2) + 10\log f - 10\log\left(\frac{\eta_{\text{net}}R_{\text{PD}}}{qF_A}\right). \tag{18}$$

In linear units, the equation is

$$P_{1\text{pump}}(B,f) = f\,\frac{E_{\text{shot}}(B)}{\eta_{\text{net}}}, \tag{19}$$

$$E_{\text{shot}}(B) \equiv 2^{3B}\left(\frac{3}{2}\right)^{3/2}\frac{qF_A}{R_{\text{PD}}}, \tag{20}$$

where we introduce a term, E_{shot}, that links power, frequency, and resolution in the shot-noise regime [44]. All of these terms have a physical limit since F_A is strictly greater than one, η_{net} is strictly less than one, and R_{PD} is strictly less than $hc/(\lambda q)$, which is 1.26 A/W at $\lambda = 1550$ nm.

Like in the thermal regime, received signal power increases with optical pump power squared, but, now, the noise component also increases with optical power. The result is a strong resolution scaling of $30\log 2 = 9.0$ dB/bit instead of $20\log 2 = 6.0$ dB/bit in analog electronics. Another notable feature of the expression for E_{shot} is that APD gain does not appear explicitly. The excess noise, F_A, increases with M meaning that APDs strictly increase the power needed to achieve a given resolution in the shot-noise regime.

3. RIN regime

RIN is due to random changes in the power output from carrier lasers. For the RIN relation, we combine Eqs. (9) and (A15):

$$20\log 2 \times B + 10\log(3/2) = \frac{2}{3}\left[-x_{\text{RIN}} - 10\log F_A + 10\log 4 - 10\log f\right]. \tag{21}$$

There is no power in this expression, so we rearrange in terms of frequency. This is the maximum frequency that can be obtained at a given bit value.

$$10\log f \le -30\log 2 \times B - \frac{30}{2}\log(3/2) - x_{\text{RIN}} - 10\log F_A + 10\log 4. \tag{22}$$

In linear units, it is

$$f \le F_{\text{RIN}}(B), \tag{23}$$

$$F_{\text{RIN}}(B) \equiv 2^{-3B}\left(\frac{2}{3}\right)^{3/2}\frac{4}{F_A}10^{-x_{\text{RIN}}/10}, \tag{24}$$

where we define a term $F_{\text{RIN}}(B)$, the maximum viable bandwidth of an APL of a given resolution. Using typical values ($M = 1$, $F_A = 1$, and RIN $= -155$ dB/Hz), that maximum bandwidth is approximately $F_{\text{RIN}}(B) \approx 2^{-3B} \times 6.9 \times 10^{15}$ Hz.

RIN imposes a hard limit on the ability of laser light to represent analog signals. This limit applies regardless of how signals are generated and detected, or how powerful the laser is. The resolution limit is 7.5 effective bits at 1 GHz and 5.3 bits at 100 GHz. Stated as a bandwidth limit: at 4 bits, the maximum frequency is 1.7 THz; at 6 bits, it is 26 GHz; at 8 bits, it is 410 MHz.

Table II calculates typical values for the metrics describing required laser power as limited by thermal, shot, and relative intensity noise. The first row pertains to a chip that is fabricated with a 50-Ω junction resistor, while

the second row uses optimally designed junction resistors whose optimum value varies depending on the operating bandwidth. E_{thrm} and E_{shot} describe situations where more power is needed to support more bandwidth. To be invariant quantities, they therefore must have units of energy, even though it is not obvious how they correspond to a physical packet of light or electricity. These quantities are useful because they can be compared directly to energies of detection and digitization that might be present around an APL. As an example, E_{shot} (physical limit) means that, for a 1-GHz system, the APL would require a minimum 0.96 μW of optical power to support a 4-bit signal resolution.

C. Multiple channels

In photonic neural networks and vector-matrix multiplication (VMM), each input channel has a corresponding optical pump and modulator. Each signal must have the potential to fan out to the different outputs, whether fan out occurs in a broadcast splitter (multiwavelength architecture) or within a MZI mesh (coherent architecture). By energy conservation, fan out carries an attenuation factor of $1/N$ [45]. Fan out in MZI architectures is revisited in more detail in Sec. IV B. At the same time, analog summation means that output signal power can recover some of this fan-out attenuation, leading to an apparent fan-in gain. Fan-in gain is dependent on the signals' cross-correlation, so we must introduce a term to quantify cases of signal correlation.

1. Fan in with correlated signals

As a result of additive fan in, the rms power of the electrical output signal depends on the values of the inputs, meaning that SFDR and resolution become signal dependent. Every situation lies somewhere on the continuum between these three cases. These cases are (worst or singular case): all received signals, after weighting, are zero except for one, (uncorrelated case): all inputs have the same rms and are statistically independent, (best or identical case): all inputs are the same. The three special cases are illustrated in Fig. 3. The effects of fan out and fan in can be stated as modifications to the total received photocurrent I_{rec} from Eq. (A1). We use $I_{\text{rec}}|_{N=1}$ to indicate the baseline value.

$$\begin{aligned}\text{rms}\,[I_{\text{rec}}(N)] &= N^{-1}\, I_{\text{rec}}|_{N=1} \quad \text{(singular case)},\\ &= N^{-1/2}\, I_{\text{rec}}|_{N=1} \quad \text{(uncorrelated case)},\\ &= 1\, I_{\text{rec}}|_{N=1} \quad \text{(identical case)},\end{aligned} \tag{25}$$

where the output signal amplitude is a statistical value given by

$$\text{rms}\,(I_{\text{rec}}) \equiv \sqrt{\left\langle (I_{\text{rec}} - \langle I_{\text{rec}}\rangle)^2 \right\rangle}. \tag{26}$$

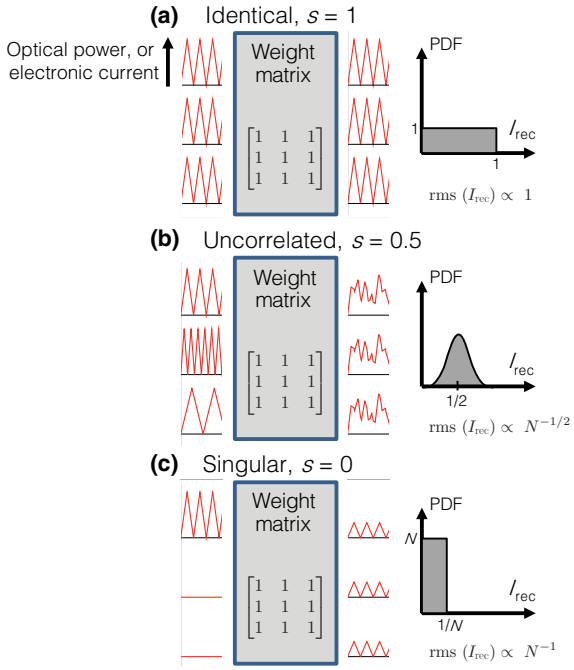

FIG. 3. Correlation factor, fan-out, and received signal power in any black-box weight matrix based on power- or current-modulated signals. Left panels show different input signal cases and resulting outputs (to scale). All weight matrices are fully transmitting. Right panels show the probability density functions (PDFs) of output voltage, the rms of which determines signal power, which is compared to noise power to obtain SFDR. (a) In the identical case, all power entering the matrix leaves the matrix, regardless of N. (b) When inputs are uncorrelated, their sum approaches a Gaussian distribution whose width (i.e., signal power) decreases with N. (c) In the worst case, only one signal is nonzero, and it must attenuate by a factor of N in order to fan out to N output ports. This behavior is a consequence of photon and charge conservation and holds regardless of weight-matrix implementation.

We can introduce a similarity variable, s, to cover these cases such that $s = 0$ is singular, $s = 0.5$ is uncorrelated, and $s = 1$ is identical [46]. In general,

$$I_{\text{rec}}(N) = N^{s-1}\, I_{\text{rec}}|_{N=1}\,, \tag{27}$$

where s can take on continuous values between 0 and 1, between the extreme cases shown in Fig. 3. The exact value of s depends on the situation, specifically, on the cross-correlation of input signals and the weight matrix. We leave its general expression for further work.

The multichannel modification to average received photocurrent has effects both on signal and noise. In all noise regimes, adding channels increases average photocurrent and therefore maximum signal amplitude. Noise amplitude can grow at the same rate or slower than the signal amplitude, depending on the type of noise.

2. Thermal regime

No matter how many channels are present, there is still only one photodetector per channel, so thermal noise does not depend on N. Correlation-dependent fan in improves only SFDR. We can calculate this effect by substituting this multichannel variant of I_{rec} into Eqs. (A4) and (A6), and so on to reach a multichannel variant of Eq. (15).

$$P_{1\text{pump}}(N,f,B) = N^{1-s}P_{1\text{pump}}(1,f,B) \quad \text{(thermal)}. \quad (28)$$

That means, to maintain the SFDR and thus resolution, in the worst case, the laser power must increase in proportion to fan out, N. In the uncorrelated signal case, it must increase by only $\sqrt{N}$.

In a $N \times N$ photonic network, total laser pump power carries an additional factor of N to provide power to all of the input channels. The total laser power required by the entire network is thus

$$P_{N\text{pumps,thrm-limit}} = N^{(2-s)} f \frac{E_{\text{thrm}}(B)}{\eta_{\text{net}}}. \quad (29)$$

The result depends significantly on the signal correlation variable, s. The result is that thermal noise-limited system power in a situation-dependent continuum somewhere between activity proportional (best case) and MAC proportional (worst case).

3. Shot regime

Shot noise depends on the total received power regardless of its wavelength, so shot noise is not independent of N. From Eq. (A6), we see that shot-noise power is proportional to I_{rec}, which means that it scales as

$$p_{\text{shot}}(N) = N^{s-1}\, p_{\text{shot}}|_{N=1}\,. \quad (30)$$

Making a similar rearrangement to get needed power for a single-channel APL and an $N \times N$ network,

$$P_{1\text{pump}}(N,f,B) = N^{1-s/2}P_{1\text{pump}}(1,f,B) \quad \text{(shot)}, \quad (31)$$

$$P_{N\text{pumps,shot-limit}} = N^{(2-s/2)} f \frac{E_{\text{shot}}}{\eta_{\text{net}}}. \quad (32)$$

Shot-noise-dominated power is therefore MAC proportional at worst. Compared to the thermal regime, the favorable effect of analog summation scales less strongly with s.

4. Coherent RIN regime

Relative intensity noise manifests in a fundamentally different way between multiwavelength and coherent architectures. Whereas thermal and shot noise have to do with detection, which is the same in each, relative intensity noise pertains to the optical domain where wavelength matters.

In order to create a coherent phase state between channels, coherent weighting architectures must have exactly one laser source. When a single laser is used, the noise in the optical domain is correlated over all channels, so noise grows at the same rate of signal amplitude. Equation (A7) is applicable. As a result, the RIN-limited SFDR stays constant with N

$$f \le F_{\text{RIN}}(B) \quad \text{(coherent)}. \quad (33)$$

5. Multiwavelength RIN regime

When multiwavelength architectures use different lasers for each channel, their intensity fluctuations are uncorrelated. The sum of uncorrelated fluctuations grows more slowly than the sum of partially correlated signals, so Eq. (A7) is *not* applicable. Take, for example, the identical case of $s = 1$: the received photocurrent is multiplied by N but p_{RIN} is multiplied by $\sqrt{N}$ because it is a sum of uncorrelated random variables.

As a result, RIN growth rate is the same as for shot noise, but the reasoning is different. Shot noise stems from photon detection, so it depends on the square root of the sum of received optical signals. RIN depends on the sum of noise across received optical signals (not the square root), but these signals each have their own independent noise components. The sum of these noise components grows only with the square root of the number of channels. For an N-to-1 fan-in circuit as well as an $N \times N$, the RIN limit becomes

$$f \le N^{s/2} F_{\text{RIN}}(B) \quad \text{(multiwavelength)}. \quad (34)$$

Relative intensity noise plays a role in capping the maximum viable frequency of an analog photonic network. This is one aspect of the differences between multiwavelength and coherent architectures. Taking, for example, a 32-channel network with 6-bit signals that are uncorrelated ($s = 0.5$), a single-source architecture would be capped to 26 GHz and multisource WDM architecture to 62 GHz. It is worth noting that multiwavelength architectures can be driven by single multiwavelength lasers [47]. In these cases, the RIN over channels is correlated, and the single-source limit is applicable.

D. Foreseeable technology

We identify waveguide-integrated APDs [48,49] and low-noise pump lasers [50] as critical technologies for neuromorphic photonics. The effect of APD on link parameters is shown in Fig. 8. APD gain pushes the thermal-limited SFDRs to the left. In other words, they reduce the pump power needed to get a certain SFDR but do not increase the maximum achievable SFDR constrained by RIN. APDs amplify the signal in the photocarrier domain, between photons and photocurrent. Once the signal becomes a

photocurrent in a circuit, thermal noise arises, after which purely electronic amplifiers can only degrade SFDR. APDs contain an extra dopant layer where electrons or holes are energetic enough to create more free pairs. In turn, these pairs create more free pairs and so on, hence the name avalanche. An APD comes with an additional noise factor, F_A, Eq. (A8), that multiplies optical noise (shot and RIN) but does not contribute to thermal noise.

For concrete values, we below consider the device from Martinez *et al.* [48] because it requires no change to the baseline silicon photonic process. Its maximum gain is $M = 60$, although its $M = 10$ operating point is more favorable overall for reasons described in Sec. VI B.

Relative intensity noise in lasers imposes a hard tradeoff between bandwidth and resolution. This RIN limit could be improved using low-noise lasers or modulator nonlinearity compensation. In Ref. [50], RIN was improved from its typical −155 dB/Hz to −160 dB/Hz, which would result in a bandwidth limit increase of 3.2× at a given resolution. RIN can instead be canceled in a balanced detection architecture [41], but this is a challenging prospect because it would require an exact copy of the entire neural-network architecture.

IV. GAIN CASCADABILITY

For a computational gate or neuron to work within a larger system, each neuron must be capable of driving the next neuron, referred to as cascadability. Many nonlinear devices can be made to exhibit input-output relationships that appear neuronlike, but this does not mean that the device can be connected with other similar devices.

Cascadability has both a physical aspect—the optical carrier properties must be compatible from input to output—and a signal amplitude aspect. A cascadable photonic neuron must be able to amplify upstream signals to drive one or more downstream neurons at an equivalent strength. This is the cascadability condition, stated as $g \geq 1$, where g is the differential optical-to-optical gain:

$$g = \left.\frac{dP_{\text{out}}}{dP_{\text{in}}}\right|_{P_{\text{out}}=P_{\text{in}}}. \tag{35}$$

We consider cascadability in a more definite way in reference to a particular circuit called the autapse: a neuron with one connection fed directly back to itself. An autapse based on a MRR modulator neuron is pictured in Fig. 4(b). It is an ideal circuit for studying cascadability because its input and output are the same by definition. Therefore, it gives insight into indefinite cascadability. We studied autapse behavior experimentally in Refs. [12,51] and, here, explore its relevance to power metrics and larger neural networks.

A. Autapse energy

The cascadability condition requires that $g \geq 1$, which leads to a minimum pump power. This pump power for modulator-class neurons was derived in Refs. [12,51], so we omit the derivation here but add in a term for APD gain, M.

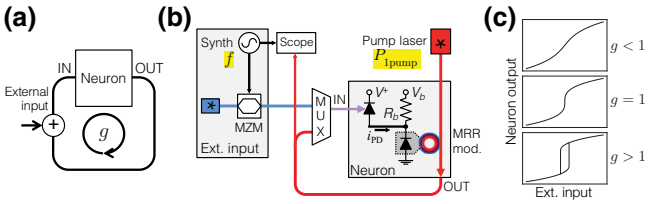

FIG. 4. A photonic autapse used to measure a device gain and bandwidth metrics experimentally. (a) An autapse is a neuron fed back to itself with a unitless round-trip gain. (b) MRR modulator neuron with a feedback connection and an external input, adapted from [51]. The modulator is pumped by a laser using $P_{1\text{pump}}$. The external input is a sinusoid or triangle wave with frequency f. (c) Cusp bifurcation in response to increasing pump power. With a weak pump, the response is monostable and nonlinear. At a certain pump power, the round-trip gain crosses 1. The transition to a bistable transfer function is evidence of this crossing. Autapse energy is measured by this experiment as $E_{\text{aut}} = P_{1\text{pump}}\big|_{g=1}/f$.

$$P_{1\text{pump}}\big|_{g=1} = \frac{2V_\pi}{\pi M R_{\text{PD}} R_b}, \tag{36}$$

where $P_{1\text{pump}}$ is the optical power entering the modulator, and R_{PD} is photodiode responsivity. V_π, called the π voltage, characterizes the modulation slope efficiency (dT/dV). A modulation voltage of V_π is needed to induce a π phase shift in a straight waveguide modulator. This shift corresponds to a 0–1 change in transmission in a Mach-Zehnder modulator. It is not the same as the voltage to induce a π shift around an MRR modulator but can be used to quantify MRR modulation slope efficiency. R_b is the impedance of the PD-modulator junction, which can be set externally. If no APD is used, then $M = 1$.

There is a problem in using Eq. (36) as a metric because R_b is a free design parameter. If the designer chooses a higher R_b, less pump power is needed, and bandwidth decreases. Bandwidth is determined by $f = (2\pi R_b C_{\text{mod}})^{-1}$, where C_{mod} is the total modulator capacitance. We can make an invariant metric by canceling this frequency

$$E_{\text{aut}} \equiv \frac{P_{1\text{pump}}\big|_{g=1}}{f}, \tag{37}$$

$$= \frac{4C_{\text{mod}}V_\pi}{MR_{\text{PD}}}, \tag{38}$$

where we call this quantity the "autapse energy," E_{aut}. This expression contains only device properties, so it is now invariant for a particular platform. Autapse energy does not correspond to any isolated operation; it has units of energy only because it is the ratio of a power to a frequency.

One favorable property of using autapse energy as a metric is that it can be measured with the same experiment used to demonstrate cascadability. With sufficient gain, the autapse undergoes a readily observable cusp bifurcation from monostable to bistable dynamics. This power is the cusp power, P_{cusp}. It was shown in Ref. [51] that the conditions for cascadability and bifurcation are the same, illustrated in Fig. 4(c). Measurement proceeds by building an autapse, increasing injected power until this qualitative transition occurs, and then measuring the power entering the system. The autapse energy is then the ratio of provided power to input signal frequency.

1. Gain in VMMs

VMMs consisting of similar photonic hardware with the addition of digital inputs and outputs have a modified gain condition. VMMs still have a unitless electrical-in to electrical-out gain, so they yield to this gain analysis. Instead of being set by cascadability ($g \geq 1$), the gain condition is set by the minimum voltage swing detectable on the input analog-to-digital converter (ADC) to the maximum voltage swing provided by the input DAC.

$$g_{\text{DA-AD}} = \frac{V_{\text{min,ADC}}}{V_{\text{max,DAC}}}, \tag{39}$$

where $g_{\text{DA-AD}}$ can serve as a performance metric of the pairing of DAC and ADC technologies. VMMs are not neural networks, so they have no cascading layers or autapses. Despite there being no autapses, the term we call autapse energy is simply a quantification of unitless gain (g) of a modulator and detector platform. That means we can define

$$E_{\text{aut,VMM}} = g_{\text{DA-AD}} \frac{P_{1\text{pump}}\big|_{g=1}}{f}, \tag{40}$$

$$= g_{\text{DA-AD}} \frac{4C_{\text{mod}}V_{\pi}}{MR_{\text{PD}}}, \tag{41}$$

where $E_{\text{aut,VMM}}$ serves the same role as E_{aut}, even though there are no explicit autapses in the VMM. Using this slightly modified VMM autapse energy, all of the below scaling analysis and above noise analysis applies to VMMs without modification.

B. Cascadability in networks

The autapse corresponds to a linear chain of optoelectronic elements, but, of course, we are interested in the performance of networks of many neurons. $N \times N$ networks also have a cascadability condition, which we can state in terms of the autapse energy. Meeting the cascadability condition is essential in networks; otherwise, activity will attenuate to zero over time (recurrent topologies), to zero over layers (feedforward topologies), or below a detectability threshold (VMM systems). These conditions are always true in systems designed to handle the worst case, although various neural-network applications have been found to require slightly less than unity cascadability within each layer [52].

Previous works have derived required laser power in networks optimistically, $P \propto N$ [9,11,13,53], including in analog electronics [54]. We argue based on energy and charge conservation that the more pessimistic relationship, $P \propto N^2$, is the only reasonable baseline to describe any implementation of general-purpose neural networks or VMMs based on optical power- or current-modulated signals.

1. All-ones matrices

When interconnected, the optical signal from a neuron fans out to N downstream neurons. To conserve energy, this always incurs a N-fold optical fan-out attenuation [45,55]. Fan-out attenuation can be countered by fan-in gain. Fan-out attenuation and fan-in gain cancel completely only in one very special and computationally trivial case shown in Fig. 3(a): a maximally transparent weight matrix of all ones ("one" being normalized against insertion loss) and identical inputs ($s = 1$). In this trivial case, network energy scales with NE_{aut}.

In order to be general purpose, neural networks and VMMs must be able to implement any weight matrix, including the all-ones matrix. They also must be able to handle any combination of input signals, including the $s = 0$ case, where fan-in gain is zero. To meet this condition, each neuron must be able to drive all downstream neurons. The cascadability condition thus becomes $g \geq N$ for every neuron, and the aggregate pump power needed to meet cascadability becomes

$$P_{N\text{pumps,gain-limit}} = N^2 f \frac{E_{\text{aut}}}{\eta_{\text{net}}}, \tag{42}$$

where η_{net} is excess insertion loss of the interconnect, described in more detail below.

Previous works have found this scaling relation to be N proportional by the following means. In Ref. [53], this was found by imposing a cap of $1/N$ on the absolute value of all weights. We argue that this is not a useful definition because the optical power available to each receiver approaches zero as N increases in all but the $s = 1$ case. Even if the cascadability condition could be relaxed in this way, thermal noise (Sec. III) would quickly come to dominate. Reference [11] proposed an $O/E/O$ neuron that countered optical attenuation with electronic gain. At low gains, amplifier power consumption is constant with gain, which results in an apparent N proportionality in the network power. Eventually, however, amplifier power consumption does scale with gain, which leads to that additional factor of N. We discuss how our analysis could

extend to electronic amplifiers in Sec. VII. Others have arrived at N proportionality through a "fixed precision" argument [9], originally outlined in Ref. [54]. In this argument, there is an embedded assumption that the sum of uniformly distributed random variables is also uniformly distributed, which is incorrect in all but the trivial $s = 1$ case, as can be seen by comparing Figs. 3(a) and 3(b). The sum of nonidentical random variables approaches a Gaussian distribution whose variance (i.e., signal amplitude) decreases with the number of inputs.

2. Unitary and permutation matrices

The quadratic scaling of gain-limited pump power applies both to multiwavelength and coherent architectures. This is apparent for the multiwavelength architecture because of the explicit fan out in their broadcast splitter, but it is less apparent for MZI mesh architectures unless considering the need to handle general cases of signal vectors and weight matrices.

MZI meshes are unitary optical devices in which the total output power can be equal to the total input power and distributed in an arbitrary way over the output waveguides. It has been argued that, therefore, all unitary matrices are also special cases for MZI VMMs, which result in N-proportional power scaling [6]. Any matrix can be made unitary by scaling singular values of the matrix, an approach taken in Ref. [56]; however, this network-level weight scaling pushes a factor of N into a network-level insertion loss. In special cases where all singular values are unity (i.e., $\Sigma = I$), the aggregate output *optical* power can be equal to the aggregate input optical power; however, the aggregate rf *signal* power decreases as a function of N. This is because the outputs are weighted sums of nonidentical signals.

In systems carrying power-modulated signals, light entering different input ports is modulated by signals from different channels. These modulations are not identical and not interchangeable in the same way that unmodulated light is indistinguishable in unitary optical devices. That means there must be a fan out of modulated energy within the mesh, even if there is not an explicit fan-out device. Every signal needs enough power to potentially reach all detectors with sufficient strength. It follows that, in order to cover the general case, each channel must provide N-proportional power, resulting in an N^2-proportional total laser power, the same scaling law that applies to multiwavelength architectures.

MZI meshes have additional special cases in the form of permutation matrices and unscrambling matrices. A MZI mesh implements a permutation by rerouting all power from input i to output j. Effectively, there is no fan-out attenuation and thus no quadratic power scaling. A promising application of MZI meshes is separating signals that have been mixed in a unitary way [24]. Power scaling is proportional to N in unscramblers because, when viewed together, the mixing media (A) and unscrambler system (A^{-1}) constitute a permutation matrix. This special—albeit useful—case does not extend to general matrix cases.

3. Insertion loss

Insertion loss is represented by the η_{net} variable. It includes waveguide propagation loss, which depends exponentially on waveguide propagation length. Propagation length is proportional to the number of neurons times either the pitch of the MRRs or length of MZIs. This means insertion loss is nondominant up to some threshold in N, at which point, it becomes strongly dominant. For MRR calculations below, we assume a 20-μm MRR pitch, a 50-μm MZI length, and 1.0-dB/cm waveguide propagation loss. MRR weight banks also incur a fixed insertion loss due to an imperfect ability to distinguish wavelengths. As determined in Refs. [57,58], a typical weight-bank insertion loss is 3 dB. Insertion loss is explicitly present as η_{net} in all relevant expressions below because waveguide loss is present in all photonic systems.

4. Sub-all-to-all networks

Sub-all-to-all networks in which there are less than $N \times N$ possible connections have numerous applications, such as for convolutional and modular networks. Other hardware architectures not studied in this paper have been developed to handle these applications, such as, respectively, in Ref. [59] and Ref. [60]. An analysis of fixed fan out was performed in Ref. [60]. In general topologies, for one neuron to be cascadable to all of its downstream neurons $g_i \geq N_{\text{FO},i}$, where $N_{\text{FO},i} \leq N$ is the fan out of neuron i. Summing over all neurons,

$$P_{N\text{pumps,gain-limit}} \propto \Sigma_{i=1}^{N} N_{\text{FO},i}. \tag{43}$$

The sum of fan out over all neurons is the same as the total number of weights in the network. Thus, Eq. (42) could be generalized to arbitrary topologies by replacing N^2 with the number of possible weights in the network.

C. Foreseeable technology

Equation (38) has the device-level expression for autapse energy. Device values are listed in Table III. We identify critical technologies as interleaved and vertical junction modulators [20,61], waveguide-integrated APDs [48], and graphene-based modulators [62]. The best modulators for silicon photonic neural networks could be different from those that are best for optical communication.

For concrete values, we take as baseline a lateral depletion modulator with switching charge of about 100 fC [40]. The interleaved junction in Ref. [61] was reported to have $C_{\text{mod}} = 3.3$ fF and approximate drive voltage of $V_\pi \approx$

TABLE III. Cascadability metrics of modulators.

Name	Variable	Value	Description
π voltage	V_π	1.5 V	Voltage to induce a half-period shift in an equivalent MZM modulator (baseline [40])
		0.5 V	Vertical junction microdisk modulator [20]
		0.95 V	Graphene push-pull modulator (proposed) [62]
Capacitance	C_{mod}	35 fF	Depletion modulator capacitance (baseline [40])
		17 fF	Vertical junction microdisk modulator [20]
		0.27 fF	Graphene push-pull (proposed) [62]
Autapse Energy	E_{aut}	220 fJ	Defined in Eq. (38) (baseline [40])
		35 fJ	Vertical junction microdisk modulator [20]
		1.1 fJ	Graphene push-pull modulator (proposed) [62]

2.2 V, resulting in charge of 7.2 fC. The vertical junction [20] has higher capacitance and lower voltage resulting in 8.5 fC. Switching charge can be further reduced using more advanced modulators incorporating alternative materials. The graphene plasmonic slot modulator studied by Ma *et al.* [62] could reduce charge to 450 aC. There is a fundamental limit on modulator switching charge set by one electron (0.16 aC), but current devices are far from this limit.

V. OPTOELECTRONIC TRANSDUCTION

In modulators and detectors, there is power dissipated in the form of current flowing through the circuit. In all photonic information processors, there is E/O conversion at the front and O/E conversion at the back in order to interface with electronic signals. A single term can combine power from the front E/O, back O/E, and any intermediate $O/E/O$ converters (a detector directly connected to a modulator). The analysis applies to both multiwavelength and coherent architectures on a given optoelectronics platform. The contribution to the network power scales with the number of circuits and their bandwidth, which we refer to as activity proportional.

$$P_{\text{OEO}} = Nf \times E_{\text{OEO}} \tag{44}$$

$$E_{\text{OEO}} = E_{\text{mod}} + E_{\text{det}} + E_{\text{ADC}}. \tag{45}$$

where P_{OEO} is the network total power, and E_{OEO} is another variable. The $O/E/O$ energy has terms associated with the modulation, detection, and ADCs if present. We find that modulator switching energy is always insignificant. Photodetector transduction energy becomes significant when using avalanche photodiodes but not when using p-i-n diodes. With larger networks, activity-proportional (Nf) contributors eventually become amortized by MAC-proportional ($N^2 f$) contributors. The amortization crossover point can be up to hundreds of neurons under certain conditions, so we cannot simply neglect this $O/E/O$ contributor.

A. Modulator charging

The dynamic energy needed to operate a depletion modulator is related to C_{mod} and V_π, where C_{mod} is its parasitic capacitance, and V_π is the voltage swing needed to modify the transmission. We focus on modulator neurons in this analysis, pictured in Fig. 1 and discuss laser neurons in Sec. VII. Average power is activity dependent. In the worst case, the modulator has to charge every other time interval, just like in an optical communication system.

$$E_{\text{mod}} = \frac{1}{4} C_{\text{mod}} V_\pi^2. \tag{46}$$

B. Photocurrent flows

Power dissipated in each PD can be derived from its I-V activity. Appendix B derives these currents, placing a constraint that the optoelectronics adhere to the cascadability condition. The result is a metric describing the energy of detection in cascadable, analog photonic signal processors. Restating Eq. (B13),

$$E_{\text{det}} = 4 V_\pi C_j V_d, \tag{47}$$

where V_d is the bias voltage on each photodetector, C_j is the combined capacitance of modulator and detector, and V_π describes modulator slope efficiency. E_{det} is favorable as a metric because it is invariant of operational variables and free design variables. It depends only on properties of the optoelectronic platform. Here, we discuss how detection energy can be compared to modulation energy and autapse energy, and the resulting implications for system design.

All detectors must be biased at high enough voltage to avoid becoming forward biased when providing their maximum photocurrent.

$$V_d > i_{\text{max}} R_b \tag{48}$$

$$> \frac{2 V_\pi}{\pi} \tag{49}$$

$$E_{\text{det}} > \frac{8}{\pi} C_j V_\pi^2. \tag{50}$$

As a consequence of the cascadability condition, modulation energy is always insignificant compared to detection energy, regardless of either technology.

$$\frac{E_{\text{det}}}{E_{\text{mod}}} > \frac{32}{\pi} \approx 10.2. \tag{51}$$

Noticing that Eq. (47) has a charge term, detection energy can also be situated in relation to autapse energy. The ratio determines which power contributor will dominate at different operating points.

$$\frac{E_{\text{det}}}{E_{\text{aut}}} = MR_{\text{PD}}V_d. \tag{52}$$

For a p-i-n detector, $M = 1$, and V_d can be close to its minimum set by the forward bias constraint. With typical values, the energy ratio is approximately 1. Since laser pump power scales quadratically with N, electrical power associated with detection will rapidly become insignificant compared to laser power. That is not the case when APDs are used.

APDs greatly increase both M and V_d. The combined effect is a reduction in autapse energy and increase in detection energy. APD bias voltages have a dependence on their photoelectric gain and various device properties, so we must refer to example devices to get concrete values. Suppose we take an example operating point from Ref. [48, Fig. 2(c)] for a 500-nm multiplication width: $M = 10$, $R_{\text{PD}} = 0.8$, $V_d = 16$ V. The detection to autapse energy ratio evaluates to approximately 128. What this means is that, for systems that use these APDs, detection energy will dominate when $N < 128$.

C. Analog-to-digital conversion

Digitization can play several roles in analog photonic information processors. In the linear category, all VMMs require digitization of all channels; field-programmable photonic arrays may require digitization of only a few [63]; and multivariate rf photonics aim to minimize the number of ADCs [64]. Nonlinear photonic neurons based on $O/E/O$ conversion can contain ADCs in order to tailor their transfer function [65], or they can can use electro-optic nonlinearities to avoid digitization altogether [8,66, 67].

Published ADC energies across the 2–6 bit range are around 1 pJ as of 2021 [68]. 1 pJ is not a theoretical limit. It is obtained in the Supplemental Code [69] by fitting an empirical envelope around demonstrated devices, a technique developed by Sundstrom *et al.* [70]. In the decade since Sundstrom *et al.* performed this analysis, ADC efficiency has improved by about an order of magnitude, and it could continue to improve in the future (or outside of published literature). Supposing this empirical envelope of 1 pJ, ADC energy could exceed optoelectronic detection energy by 2 orders of magnitude, so it cannot be neglected. Both can be amortized in large networks, although that amortization happens much sooner in analog systems with less than complete digitization compared to photonic VMMs requiring ADCs on every channel.

D. Foreseeable technology

There is not much room for p-i-n detectors to improve to their physical limits, and $O/E/O$ power is already non-dominant compared to contributors to laser power. The introduction of APDs, however, can substantially reduce laser power, at the same time that their higher bias voltage rapidly increase $O/E/O$ power. If APDs are used, there is a balance to be struck between APD gain, APD bias voltage, and autapse energy, such that both optimal device design and biasing point depend on operating conditions (N, f).

VI. SUMMARY AND ROADMAP

We analyze the silicon photonic neural-network architectures of Figs. 1(a) and 1(d) when implemented on a modern-day mainstream silicon photonic foundry platform [40]. There are four key results of this exercise: (1) an overall power equation, (2) the derivation of metric quantities describing performance, (3) the discovery of regimelike behavior in scaling behavior, and (4) a roadmap of key foreseeable technologies and their quantitative significance. The choice of roadmap technologies is constrained to devices that have been demonstrated in a research setting. The rationale for these choices is based partly in addressing limiting power components (see Sec. VI B) and partly in feasibility to incorporate into silicon photonic foundry processes. In all regimes, power scales superlinearly with the number of neurons, with the exception of the $O/E/O$ regime. This finding significantly refines and, in some sense, conflicts with many presumed scaling laws for neuromorphic photonics [7, Sec. 14.2.3] and [6,11–13].

A. Overall power equation

The power components of a neural network are shown in Table IV. The last column gives estimated coefficient values for a baseline foundry platform. In parentheses are the coefficient values given foreseeable technology currently in the research phase. The total power is the sum of each proportionality term (column 2) multiplied by the coefficient (column 5). Henceforth, we refer to this sum as the overall power equation. A glossary of all variables used is found in the Supplemental Material [17].

Power is used by lasers, weight configuration, and electrical currents flowing in modulators and detectors.

$$P = N^2 \times P_{\text{wei}} + N \times P_{\text{1pump}} + Nf \times E_{\text{OEO}}. \tag{53}$$

TABLE IV. Metrics for silicon photonic neural networks.

Component	Prop. to	Coefficient	Equation	Value (Foreseeable)
Weight locking	N^2	MRR: $P_{\text{lock}} = K\Omega(Nd)$; MZI: N/A	(4)	14 mW (74 nW)
Weight configuration	N^2	MRR: $P_{\text{conf}} = K/2\mathcal{F}$	(5)	120 μW (230 nW)
		MZI: $P_{\text{conf}} = 2P_{\pi,\text{MZI}}$	(6)	20 mW (200 nW)
Pumping: thermal limit ($B = 4$)	$N^{(2-s)}f$	$E_{\text{thrm}} = 2^{(3/2)B}\,(3/2)^{3/4}\,\sqrt{8\pi k_B T}\left(\sqrt{C_{\text{pd}}}/MR_{\text{PD}}\right)$	(29)	6.5 fJ (58 aJ: with APD $M = 10$)
Pumping: shot limit ($B = 4$)	$N^{(2-s/2)}f$	$E_{\text{shot}} = 2^{3B}\,(3/2)^{3/2}\,(qF_A/R_{\text{PD}})$	(32)	1.5 fJ (960 aJ: uncompensated, 98 aJ: modulator nonlinear compensation, Sec. VII D)
Pumping: gain limit	$N^2 f$	$E_{\text{aut}} = (4V_\pi C_{\text{mod}}/MR_{\text{PD}})$	(42)	260 fJ (128 aJ: with APD $M = 10$)
$O/E/O$	Nf	$E_{\text{OEO}} = \frac{1}{2}C_{\text{mod}}V_\pi^2 + 4C_j V_\pi V_d + E_{\text{ADC}}$	(45)	220 fJ (680 aJ)
Bandwidth: RIN limit	$N^{s/2}$	$F_{\text{RIN}}(B) = 2^{-3B}\,(2/3)^{3/2}\,(4/F_A)10^{-\text{RIN}/10}$	(24)	1.7 THz (5.3 THz)

The power needed for each weight down into static locking power and the power needed to configure the weight values: $P_{\text{wei}} = P_{\text{lock}} + P_{\text{conf}}$, which are expressed as

$$N^2 \times P_{\text{wei}} = N^2 \times K\Omega(Nd) + N^2\frac{K}{2\mathcal{F}} \quad \text{(multiwavelength)}, \tag{54}$$

$$N^2 \times P_{\text{wei}} = N^2 \times 2P_\pi \quad \text{(coherent)}, \tag{55}$$

where K is tuning efficiency in mW/FSR, Ω is the expected amount of tuning per weight (in FSRs) needed to counteract fabrication variation, d is MRR weight pitch, and $\mathcal{F}$ is finesse.

The power needed for laser pumps is determined either by the cascadability threshold or the signal resolution.

$$N \times P_{\text{1pump}} = N^2 f \times \eta_{\text{net}}^{-1} \max\left[E_{\text{aut}},\, N^{-s}E_{\text{thrm}},\, N^{-s/2}E_{\text{shot}}\right], \tag{56}$$

where E_{aut} is the autapse energy, η_{net} is network transmission efficiency—dependent on Nd—s is a signal correlation factor from 0 to 1, E_{thrm} is a thermal-dominated resolution coefficient defined in Eq. (15), and E_{shot} is a shot-dominated resolution coefficient defined in Eq. (20). The last $O/E/O$ term, is due almost entirely to current flowing in photodetectors and ADCs.

$$Nf \times E_{\text{OEO}} = Nf \times (E_{\text{mod}} + E_{\text{det}} + E_{\text{ADC}})\,. \tag{57}$$

The overall power equation exhibits an intricate, regimelike structure. This can be viewed as a result of overall power being a sum of polynomials. Regimelike behavior means that a rich vocabulary of quantitative concepts are required to discuss power use aspects and the impacts of foreseeable technology on performance; this discussion cannot be boiled down to one concept of operation-based efficiency. We next discuss some of these aspects and must rely heavily on the terms in Eqs. (53)–(56).

B. Operating regimes

Over the operational domain (scale and bandwidth), different power components dominate, leading to operating regimes. In the above sections, we discuss technological effects on individual components, but total power cannot be improved by focusing all efforts on any one power contributor. Here, we discuss total power, the interplay of regimes, and the thought process behind technology roadmapping.

Figures 5–7 plot system power and net MAC efficiency and can be reproduced by the published Supplemental Code [69]. Each panel represents a single future technology scenario, where some permutation of foreseeable technologies is incorporated. In the terminology, a "region" is a part of the domain surrounding a (N,f) coordinate. A "regime" is the entire subdomain dominated by a particular component. A region always exists at the same spot and can fall in different regimes. So for example, we can refer to the lower part of the plots as the "small-scale region."

In the plots, regimes are marked by different colors, as listed in the legend on the bottom. We can refer to all blue areas and the "gain-dominated regime" or "gain-limited regime." Regimes can move panel to panel. When moving between panels, if a regime shrinks, it means some technology has improved that contributor relative to other contributors. Next, we discuss specific roadmaps for the multiwavelength architecture.

C. Technology roadmap

The baseline state of current silicon photonic technology [40] is shown in Fig. 5(a). There is a curved boundary between weight locking and gain pumping going through $(N,f) = (1, 2\text{ GHz})$ and (800, 20 GHz). This curvature is due to the distance dependence of MRR variability in $\Omega(Nd)$. Failing to account for covariation in small networks faster than 1 GHz results in an order-of-magnitude power overestimate. There is another curved

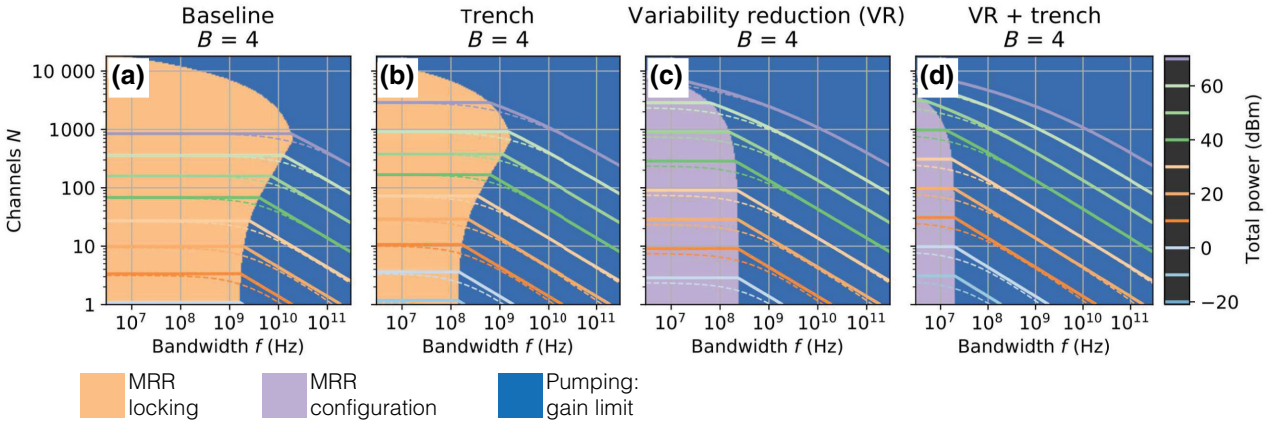

FIG. 5. Possible roadmap for MRR tuning. Colors represent dominant power contributors. Solid contours represent power due to the dominant component. Dotted contours represent the sum of all components. (a) Current-day baseline for a multiwavelength architecture. (b) Greater tuning efficiency due to trench heaters has an effect, but locking (orange) still dominates. (c) Reducing MRR variability greatly reduces locking power, but then configuration (purple) takes over. (d) The impacts of both technologies compound, giving a tuning power improvement of 3 orders; however, tuning-dominated regimes also shrink. They have no impact on required laser pump power (blue).

boundary appearing above $N = 800$. This curvature is due to a ballooning waveguide loss, $\eta_{\text{net}}(Nd)$, in networks of large physical size that must be countered by more pump power. Currently, weight locking power dominates a sizable region.

The rest of Fig. 5 shows the impacts of tuning-related technology. Trench-etched thermal tuning [Fig. 5(b)] moves the orange-blue boundary to the left by one order and power contours in the orange regime up by one order. When introducing a form of MRR variability reduction, such as trimming [Fig. 5(c)], weight locking becomes nondominant and weight tuning (purple) is dominant. In small-scale, low-bandwidth regions, trimming reduces control power by one order; in larger-scale, low-bandwidth regions, it reduces power by 2 orders. Combining trimming and trenches [Fig. 5(d)] compounds the effects.

Overfocusing on tuning-related technologies has shrunk the weight-dominated regimes to almost nothing. Over most of the domain, gain-limited laser pump power (blue) is now dominant. In other words, we can say that the combination of two tuning technologies has created an imbalance of regimes, where the improvements manifest only in small pockets. Next, we address that large, blue, now-limiting regime with modulator technologies.

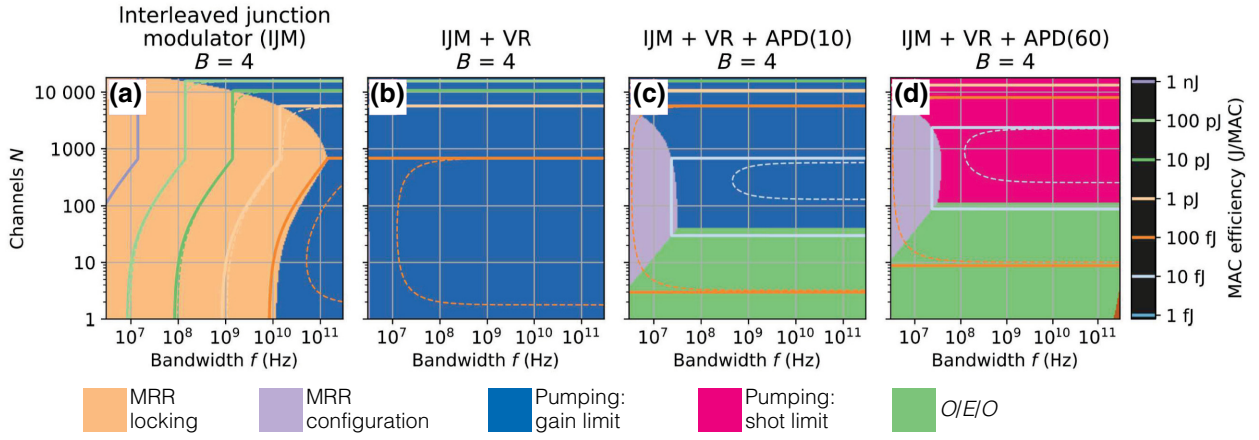

FIG. 6. Semiconductor optoelectronics roadmap. Contours represent net MAC efficiency, the ratio of total power to $N^2 f$. Solid are due to the dominant component alone, and dashed use the total power. (a) Alone, interleaved junction modulators address a nondominant regime. (b) Combined with variability reduction, the modulators have a large effect both on weight-tuning efficiency and the gain-dominated laser power. (c) Avalanche photodetectors improve the gain limit (blue) but increase $O/E/O$ powers (green) so that the $O/E/O$ contributor dominates below 50 neurons. (d) Increasing APD gain further brings minimal return because the shot-noise limit (pink) appears.

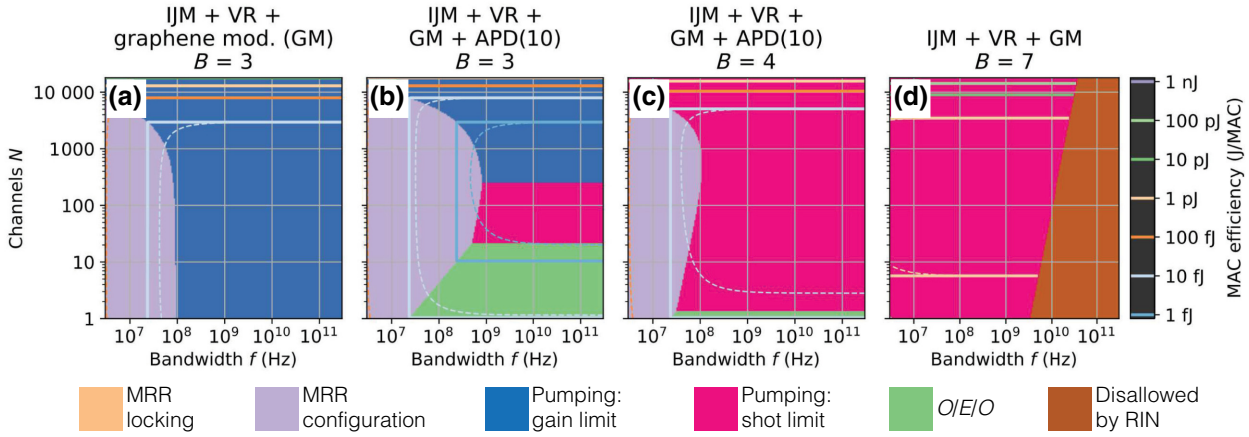

FIG. 7. Effects of noise. (a) When combined with prior technologies, low-charge graphene modulators have a substantial impact on the gain limit. (b) Graphene and APDs affect different contributors and therefore compound, opening up a sub-fJ/MAC region. (c) At a higher resolution, shot noise rapidly takes over most of the domain, bringing MAC efficiency above 1 fJ everywhere. (d) At a much higher resolution, shot noise raises the minimum MAC energy above 1 pJ across the domain. The hard frequency limit imposed by RIN and modulator nonlinearity appears.

Figure 6 shows several effects of optoelectronics affecting gain and noise: interleaved junction modulators (IJMs) and APDs. Introducing the modulator in Fig. 6(a) has a significant impact on E_{aut}; however, it shrinks the blue regime, and MRR locking dominates nearly the whole domain. The modulator alone—just like the tuning technologies alone—has created an imbalance of regimes, except in the opposite direction.

Complementary technologies can be introduced to simultaneously address multiple power contributors. Figure 6(b) combines variability reduction with interleaved junction modulator neurons. Compare it to just trimming in Fig. 5(c). This combination of tuning and modulator technologies has maintained a balance between P_{conf} and $E_{\text{aut}}/\eta_{\text{net}}$ regimes. As a result, efficiency improves across the whole domain.

More than being complementary technologies, the combination of IJM + VR enables nonthermal tuning, as described in Sec. II D. Nonthermal tuning partially explains the immense improvement of Fig. 6(b) and the disappearance of tuning-dominated regimes. At this point, the whole domain is dominated by laser pump power. This component can be improved by APDs.

Introducing an APD reduces E_{aut}, so it improves the efficiency of the gain-dominated regime in Fig. 6(c). With $M = 10$, E_{aut} decreases by a factor of 10. This can be seen by comparing the horizontal, orange contour in Fig. 6(b) to the horizontal, light blue contour in Fig. 6(c)—one order of magnitude more efficient. As a trade-off, the APD increases E_{OEO}, so the $O/E/O$ regime (green) now appears in the small-scale, high-bandwidth region. The horizontal green-blue boundary around $N = 50$ is indicative of a constant ratio between E_{aut} and E_{OEO}, which is derived in Eq. (52). Taking stock, we can see that these three complementary technologies have given about 3 orders-of-magnitude efficiency improvement across much of the domain [for example, compare the MAC contours at the $(N = 100, f = 10^9)$ point between Figs. 6(a) and 6(c)].

Figure 6(d) shows the effect of too much APD gain. On top of their effects on gain and detection energies, APDs affect shot noise negatively. As a result, the contours have barely moved compared to Fig. 6(c). What this means is that laser power must maintain at about the same level, but the reason has shifted from providing sufficient gain to providing sufficient signal-to-noise ratio. Indefinitely increasing APD gain is a dead end. In reality, APD design entails a delicate balance of three power contributors.

Figure 7 illustrates scenarios with different required resolutions. It also introduces extremely low-charge (graphene) modulators. Low-charge modulators have a small impact compared to previously mentioned technologies (VR + IJM + APD). Once those technologies are introduced though, the low-charge modulators complement them, pushing efficiency up across the whole domain. To see the modulator effect alone, compare Figs. 7(a) and 7(b) to Figs. 6(b) and 6(c). A subfemtojoule region (deep blue contour) can be seen in Fig. 7(b), although this region closes up above 3 bits of resolution.

Increasing resolution slightly in Fig. 7(c) results in a take over by shot noise, despite the presence of all the foreseeable technologies. Figure 7(d) shows a much higher 7-bit resolution. Regardless of which technologies are present, shot noise imposes a fundamental limit of 490 fJ at 7 bits, which is reflected in the contour lines of MAC energy. Furthermore, at this resolution, the RIN regime appears. RIN is not a power contributor per se, but rather a disallowed regime in terms of frequency where no amount of laser

power can provide 7 bits. The RIN boundary is slanted because signals are assumed to be uncorrelated ($s = 0.5$).

D. Relevance of MACs-per-Joule metrics

Multiply accumulate operations and synaptic operations (SOPs) are primary metrics for evaluating electronic neural networks and VMMs. In systems based on electronic gates, MAC efficiency values are usually invariant with system scale and bandwidth, and this invariance makes them good metrics. In photonics, MAC efficiency can vary substantially across different regimes of operation. That means MAC efficiency can be an unstable lens for evaluating the performance merit of photonic systems in general. At a given scale and bandwidth, net operational efficiency can be obtained by dividing total power by $N^2 f$. For example, consider the green regime in Fig. 6(c). In this activity-proportional $O/E/O$ regime,

$$P_{\text{sys}} \approx NfE_{\text{OEO}}, \tag{58}$$

$$E_{\text{MAC,sys}} \approx \frac{E_{\text{OEO}}}{N}. \tag{59}$$

This MAC-per-Joule expression is a poor metric for describing technology merit because it varies strongly with system scale. Using Eq. (59), one might think that E_{MAC} could be made arbitrarily small by imagining an arbitrarily large N. This type of extrapolation does not work because that imagined system would eventually enter a different operating regime.

MAC efficiency metrics do make sense within regimes in which power scales with $N^2 f$, in other words, in proportion to the rate of operations being performed. These include the gain-dominated (blue) and shot noise-dominated (pink) regimes. Visually, in Figs. 6–7, MAC invariance means that contour lines are spaced far apart in those regimes. The MAC invariance breaks down again for very large N due to propagation loss. Given our assumed waveguide loss value, that breakdown occurs around $N \geq 5000$.

Shot noise imposes a physical limit on MAC efficiency. All of the variables in Eq. (32) have physical limits. Because light is quantized in photon number, there is a minimum number of photons needed to represent a value, regardless of how that signal is generated, detected, or propagated through any transmission element. With perfect control of quantum number, there must be at least 2^B quanta to represent a B-bit signal. Photon detection is stochastic in classical light, and that results in a stronger noise-based scaling relation of 2^{2B}. In addition, E/O conversion applies a nonlinear transfer function that rapidly degrades SFDR, resulting in a distortion-based scaling relationship of 2^{3B}, contained in the E_{shot} metric [Eq. (20)]. At 2-bit signal resolution, the distortion-based limit is 15 aJ. At 8 bits, that limit increases by 5 orders of magnitude, becoming 10 times the published MAC efficiency of a 3-year-old, 8-bit electronic counterpart [2]. This would seem to indicate that photonic VMMs cannot compete with electronic VMMs at or above 8 bits—at least not in terms of MAC efficiency—however, distortion compensation can be employed to push towards the noise-based classical physical limit (Sec. VII D).

VII. FURTHER DIRECTIONS

We attempt to be exhaustive in identifying, exploring, and putting numbers to the factors affecting power use in MRR and MZI neural networks and VMMs. In the interest of length, we omit some key considerations and potential technologies in addition to other metrics. The key technologies we identify cannot be exhaustive because alternative technologies are arising. It is worthwhile to note that power-related technologies may be less critical than those that affect other metrics (e.g., electronic-photonic copackaging [71], MRR weight-bank channel limits [58], hybridized WDM-electronic fan-in limits [59]). Here, we summarize these omissions and discuss how the analyses developed can be extended to other photonic architectures.

A. Power considerations omitted

We do not consider wall-plug power, instead focusing on fundamental sources of on-chip energy dissipation that are needed for a photonic neural network to operate. In mobile applications, "wall-plug" power is a critical consideration. Furthermore, we do not make a distinction between overall power consumption and on-chip heat dissipation density. This distinction is crucial for heat sensitive environments (e.g., cryostats and datacenters) because the bulk of waste heat is dissipated by fiber-coupled pump lasers located outside of the temperature sensitive region.

To calculate wall-plug power, the efficiency of lasers, tuning sources, and biasing sources must be incorporated. In addition, there will be power use associated with CPU and/or lookup table calculations to determine tuning currents based on desired weights. Weight calculations could become significant when reconfiguration happens at high speed, but not at low speed. Finally, there will be cooling costs that increase with heat dissipation density.

Dissipation density is tied to device footprints. Using a MRR weight pitch of 20 μm, there is room for a $N = 500$ all-to-all network on a 1×1 cm die. Today, that die at 10 GHz would consume a kilowatt. In the foreseeable future, it could consume a watt. A 12" wafer has room for 13-k neurons in an all-to-all configuration. Of course, a wafer-scale 10-GHz all-to-all network would consume an absurd 10 s of kilowatts and run into time-of-flight issues, which is why sub-all-to-all topology ideas will be indispensable.

Data movement versus processing patterns are a critical consideration for every hardware accelerator, such as GPUs. It can be expected that data movement will be a significant power contributor in some cases. For general-purpose VMMs, the ratio of read and write to neural-network operations is high: each input vector must be loaded from caches on the printed circuit board (PCB), CPU memory, or, at worst, the Internet. On the other hand, it is worth noting that recurrent neural networks and many-layered machine learning models have a low ratio of read and write to weighed addition operations.

We neglect some prospective technologies related to power. For example, we do not consider the potential of integrated electronic amplifiers, which were analyzed in Ref. [14]. Electronic amplifiers have a gain-bandwidth-power trade-off, so they will effectively modify E_{aut} with a penalty, without significantly changing the form of the results. Optical amplifiers do not have the same trade-off, so they would be harder to incorporate in this analysis. Similarly, replacing modulator neurons with laser neurons would have a profound impact on autapse performance, perhaps resulting in an autapse power, rather than an autapse energy. Waveguides with substantially lower loss, such as silicon nitride waveguides, would directly impact the scale of feasible networks, pushing up the contours at the top edge of each plot.

Technologies for decreasing autapse energy (i.e., neuron gain) will also play a central role. We consider low-charge modulators and then APDs. APDs are favorable because they require no fabrication process modification [48]. The downside of APDs is that they increase shot noise, which imposes a physical limit on MAC energy. In cases where shot noise dominates, it would make sense to use purely electronic amplifiers [11,14], then folding their gain-bandwidth-energy contribution into an increased autapse energy.

B. Extension to other photonic architectures

Parts of the developed analysis could map to other multichannel photonic processors for which these analyses are currently absent. Integrated neuromorphic photonic architectures have taken forms besides the MRR and MZI mesh approaches in Fig. 1. Deep, layered networks yield to our analytical approach because they can be analyzed as concatenated all-to-all networks, even when the number of inputs is different from the number of outputs of each layer. Bangari *et al.* [59] proposed a hybridization of photonic and electronic fan in, which would require more detailed analysis of the receiver junction. The WDM architectures from Feldmann *et al.* [34] and the authors [8] use a broadcast loop or bus layout, as opposed to the broadcast tree from [12], which is used here. Bus and loop architectures put all weight banks on a single waveguide, so the waveguide propagation loss and η_{net} would scale with N^2 instead of N.

Sub-all-to-all topologies are subsets of the all-to-all topology and would yield to the above analyses. Sparse topologies, on the other hand, may break some of its assumptions. For example, in the brain and in the photonic neural networks considered by Shainline *et al.* [72], average fan out tends to 10^4 despite there being billions of neurons. Interconnectivity eventually scales with $1/N$, meaning a quadratic power-to-neurons proportionality will break down. Even into the future, we hazard to guess that some kind of power-to-weight proportionality, Eq. (43), will hold for sparsely connected networks.

A variety of non-neuromorphic, multichannel photonic processing architectures are also demonstrated [16]. Every photonic processor has input modulators and output detectors meaning the transduction analysis in Sec. V is applicable. The analysis of tuning power in Sec. II is applicable to programmable photonics in general, including recirculating meshes [63,73], unscramblers [24], quantum optical neural networks [74], and others. The analysis of resolution during correlated signal fan in in Secs. III are extensible to microwave photonics, including field-programmable photonic arrays [63,73] and multivariate photonics [64,75]. We postulate that most of these systems will exhibit regimelike scaling features that could benefit at least from visualization strategies in Sec. VI B and the Supplemental Code [69].

C. Signal resolution for applications

Signal resolution has a qualitative impact on the repertoire of neural networks. We do not posit what signal resolution in the $\leq$ 8-bit range would be useful for various applications [76]. Reduced-precision machine learning has shown promising results, for example, Ref. [77] showed that accuracy on the ResNet benchmark decreased by only 0.27% when using 2-bit signals (also known as activations) and 2-bit weights. Outside of ML, resolution has been debated by the neuromorphic electronics community. The IBM TrueNorth chip uses 4-bit weights and 1-bit signals [43]. The BrainScales HICANN-X [78] chip use 6-bit weights and 5-bit signals. Pfiel *et al.* [79] demonstrated usefulness of 4-bit weights in a wide range of benchmarks and postulated that additional resolution provided little return for neuromorphic processing. For the photonic DEAP architecture in Ref. [59], a sharp jump in accuracy—from failing to optimal accuracy—was simulated to occur between 2- and 4-bit signals when applied to the MNIST dataset. Recent work has shown that certain neural-network applications can be performed at dynamic resolution [52] with a corresponding reduction in noise-dominated optical power.

We do not consider spiking photonic neurons [66,80], which have substantial differences in terms of

cascadability, noise, and optoelectronics. To further complicate scaling analysis for spiking neurons, only those based on quiescently subthreshold sources have activity-dependent power dissipation [81] while others based on excitable dynamics are dominated by the constant power holding them in a ready-to-fire state [82–84]. Finally, spikes can encode information in either rate or timing. Rate codes are continuous valued, but any tasks based on temporal coding lie completely outside of the repertoire of continuous-time photonic neurons [85]. The repertoire mismatch can undermine any kind of meaningful, quantitative comparison between the two types of architectures.

D. Modulator nonlinearity compensation

Optical modulators have nonlinear E/O transfer functions that corrupt the representation of an input signal and significantly affect the resolution analysis. We define resolution based on SFDR because SFDR analysis accounts for corruptions due to this nonlinear distortion along with corruptions due to noise. Modulator nonlinearity can be partially compensated in the optical domain [86] or by predistorting modulator drive signals [87], and these techniques would necessitate a different resolution analysis. With perfect compensation, the detected current in a single-channel link could represent the input accurately over the full range of 0 to $2I_{\mathrm{rec}}$, resulting in received signal power of $p_s = R_b I_{\mathrm{rec}}^2/2$. Effective number of bits, B, can be redefined as in Eq. (9), replacing SFDR in Eq. (A13) with $\mathrm{SNR} = p_s/p_n$. Following the derivations of Appendix A and Sec. III yields

$$E_{\mathrm{thrm,NL\ comp.}}(B) = 2^B\sqrt{24\pi k_B T}\frac{\sqrt{C_{\mathrm{pd}}}}{MR_{\mathrm{PD}}}, \tag{60}$$

$$E_{\mathrm{shot,NL\ comp.}}(B) = 2^{2B}\frac{3qF_A}{R_{\mathrm{PD}}}, \tag{61}$$

$$F_{\mathrm{RIN,NL\ comp.}}(B) = 2^{-2B}\frac{16}{3F_A}10^{-x_{\mathrm{RIN}}/10}. \tag{62}$$

Perfect nonlinear compensation changes the exponent of 2^B in these expressions, causing a major impact on high-resolution systems. For example, E_{shot} would have a physical limit of 98 aJ at 4 bits, compared to 960 aJ, uncompensated (Table IV). At 8 bits, the difference is 25 fJ compared to 3.9 pJ, uncompensated. Practical compensation techniques set the value somewhere between these bounds and will be a key technology for high-resolution systems, which motivate further analytical study.

VIII. CONCLUSION

We conduct a study of power use and efficiency of fully specified architectures for silicon photonic neural networks. We find that overall power use takes a regimelike structure, which means that no less than seven metrics are needed to describe the energy efficiency merit of these processors in full. Analytical expressions for these metrics are derived and then evaluated using concrete values corresponding to demonstrated technologies. This quantitative deconstruction of power use factors has the potential to guide research and expectation in the field of neuromorphic silicon photonics.

The scaling laws governing regimes dominated by laser pump power have finite values for MAC efficiency. These N^2f -proportional laws conflict with some versions of current projections and are comparatively pessimistic to those projections. Despite what might be seen as pessimistic performance predictions, we find that photonic neural networks and VMMs can be highly competitive in terms of MAC energy compared to state-of-the-art electronics, but only under certain operating conditions. In particular, power use increases extremely quickly with signal resolution.

Scaling analysis links present performance values to future values enabled by future technology. We present a roadmap identifying key technologies that will have critical impacts on total power: postfabrication trimming [28], vertical depletion junctions [20], MEMS tuning [38], waveguide-integrated avalanche photodetectors [48], and graphene push-pull modulators [62]. These are specific, nonspeculative technologies that have been demonstrated in research settings. Thermal tuning is the most pressing problem today, and we propose a path for radical improvement for MRR-based architectures. The need to provide a certain signal gain is found to be limiting in large swaths of the future operational domain, yet it has much room to improve with foreseeable modulator technologies.

This study reveals by example that indeed alternative concepts and metrics will be needed to discuss the power of physics-based information processors. Instead of one number, MAC efficiency, there are alternative metric concepts that must be accounted for. The high level approach of starting from a standard all-to-all network, as opposed to starting from a device or operation, is predicted to bear fruit in the quantification of physics-based neuromorphic processors in general. In these other emerging physical platforms, we anticipate a similar yielding of physical invariants to efficiency regimes, although the nature of these regimes will undoubtedly be entirely different.

ACKNOWLEDGMENT

We acknowledge the foundational role of discussions with Dr. Bhavin Shastri, Dr. Mitch Nahmias, Dr. Thomas Ferreira de Lima, and Dr. Paul Prucnal. We acknowledge Dr. Richard Mirin and Dr. Sae Woo Nam for supervision. We thank Dr. Philip Yechi Ma for edits and suggestions on APL noise sections. For editorial contributions, we thank Dr. Jeff Shainline, Dr. Nikolai Klimov, Dr. Sonia Buckley,

and Dr. Norman Sanford. This work is supported by the National Resource Council (NRC) postdoctoral fellowship program.

APPENDIX A: ANALOG PHOTONIC LINK DERIVATION

In this section, we rederive the analysis of analog photonic links following the theory presented well by Marpaung [41]. An APL consists of an input rf signal, a modulator, a transmission path, and a detector. An understanding of the basic APL is foundational for arriving at resolution-frequency-power-channel relations in neuromorphic and multivariate photonics.

The goal of this analysis is to compare analog-to-digital resolutions. Digital operations have an integer number of bits that determines the ratio of maximum to minimum values that can be represented. Analog signals have a similar idea of the ratio of large-signal amplitude to the smallest signal that can be resolved from noise. When signals are large enough, saturating nonlinearities in the components cause corrupting distortions.

Spurious-free dynamic range is one such measure combining large-signal distortion limits and small-signal noise limits. SFDR can be converted to an equivalent bit resolution according to Eq. (9).

1. APL saturation

In our rederivation, in contrast to Marpaung's, we find it fruitful to start from the track of average received photocurrent, although all below results are verified to reproduce the results of Ref. [41]. The average photocurrent at the detector is

$$I_{\text{rec}} = \frac{1}{2} M \eta_{\text{net}} R_{\text{PD}} P_{\text{pump}}, \tag{A1}$$

where P_{pump} is laser power into the modulator. η_{net} is the passive attenuation of the photonic link with maximum of 1. M is avalanche gain in an avalanche photodiode. The responsivity as defined R_{PD} takes into account the device quantum efficiency only, whereas the real responsivity of an APD is MR_{PD}. For a regular p-i-n photodiode, $M = 1$.

The rf power gain of the link [41, Eq. 2.25] can be restated in terms of average photocurrent.

$$g_{\text{rf}} = \left(\frac{\pi R_b M \eta_{\text{net}} R_{\text{PD}} P_{\text{pump}}}{4 V_\pi} \right)^2 \tag{A2}$$

$$= \left(\frac{\pi R_b I_{\text{rec}}}{2 V_\pi} \right)^2, \tag{A3}$$

where R_b is the impedance of the receiver (taken here to be 50 Ω), and V_π is the voltage needed to make a Mach-Zehnder modulator (MZM) go from fully transmitting to fully blocking. V_π can characterize modulation slope efficiency in any voltage modulator, not just MZMs.

Dynamic range involves the maximum signal that can be represented without distortion. Distortion arises from the electro-optic modulator transfer function, which has a saturating cubic nonlinearity. It is characterized by the output intercept point of the third harmonic (OIP3), which is the power at which the rf power in the cubic distortion term equals the power in the linear term. From Ref. [41, Eq 2.82],

$$p_{\text{OIP3}_{\text{lin}}}[\text{W}] = R_b I_{\text{rec}}^2. \tag{A4}$$

This should be surprising because there is no dependence on the shape of the modulator parameters, unlike for gain. A steeper slope of the modulator increases gain but it also increases the amount of saturation.

2. APL noise

There are three sources of noise: thermal, shot, and relative intensity noise. Their power in linear watts are

$$\frac{p_{\text{thrm}}}{f} = k_B T, \tag{A5}$$

$$\frac{p_{\text{shot}}}{f} = \frac{q R_b M F_A}{2} \left(I_{\text{rec}} + I_d \right), \tag{A6}$$

$$\frac{p_{\text{RIN}}}{f} = 10^{\frac{x_{\text{RIN}}}{10}} \frac{R_b F_A}{4} I_{\text{rec}}^2, \tag{A7}$$

where x_{RIN} is a laser parameter with a typical value of −155 dB/Hz. I_d is dark current, usually less than received photocurrent. M is avalanche gain. Note that these are different definitions than the ones Marpaung uses by a factor of 2: when adding up noise contributions under an assumption of a lossy, matched receiver, half of the noise power goes into the matching impedance rather than the load. The Fano factor, F_A, describes noise due to avalanche gain in an APD. F_A is defined

$$F_A = K_A M + (1 - K_A)(2 - M^{-1}), \tag{A8}$$

where K_A is the carrier ionization ratio. A typical value is of K_A is 0.1 [49], so a typical value of F_A for $M = 10$ is 2.7. Avalanche noise multiplies optical sources of noise (shot and RIN) but has no effect on thermal noise. Total noise is

$$p_N[\text{W}] = p_{\text{thrm}} + p_{\text{shot}} + p_{\text{RIN}}, \tag{A9}$$

$$P_N[\text{dBm}] = 10 \log \frac{p_N}{10^{-3}}, \tag{A10}$$

where capital P_N represents the same concept except in log units. All noise terms are proportional to signal bandwidth, f , so they are often stated in log units as dBm/Hz. P_N can be stated at a particular frequency or independent of frequency, in which case it is called "power spectral

density."

$$\underbrace{P_N(f\,[\mathrm{Hz}])}_{\text{power}}\,[\mathrm{dBm}] = \underbrace{P_N[\mathrm{dBm/Hz}]}_{\substack{\text{power}\\ \text{spectral density}}} + 10\log(f\,[\mathrm{Hz}]). \tag{A11}$$

Finally, the SFDR can be calculated. Combining the noise figure, F_N, and SFDR definitions found in Marpaung [41, Eqs. (2.48) and (2.85)]:

$$\begin{aligned}\mathrm{NF} &= P_N - G + P_{\mathrm{thrm}}\\ \mathrm{SFDR} &= \frac{2}{3}\left(P_{\mathrm{OIP3}} - F_N - G + P_{\mathrm{thrm}}\right),\end{aligned} \tag{A12}$$

we arrive at

$$\mathrm{SFDR} = \frac{2}{3}\left(P_{\mathrm{OIP3}} - P_N\right), \tag{A13}$$

where OIP3 is now in log units. Equation (A13) is plotted in Fig. 8. As desired, the concept of a ratio between highest-power (determined by modulation saturation) and lowest-power (determined by detection noise) is now apparent. The SFDR can be stated at a given frequency and also independent of frequency, just like the noise power. SFDR has a fractional term, and so too does its frequency dependence:

$$\underbrace{\mathrm{SFDR}(f\,[\mathrm{Hz}])}_{\text{power ratio}}\,[\mathrm{dB}] = \underbrace{\mathrm{SFDR}[\mathrm{dB\ Hz^{2/3}}]}_{\substack{\text{power ratio}\\ \text{spectral density}}} - \frac{2}{3}10\log f\,. \tag{A14}$$

3. Operating regimes

Noise power is a logarithm of a sum, so the largest term is overemphasized—this is the origin of noise "regimes," operating points where other sources of noise are not significant. At high powers, RIN dominates as it scales with I_{rec}^2—the same rate as gain. Here, $P_N \approx P_{\mathrm{RIN}}$, so

$$\begin{aligned}\mathrm{SFDR}[\mathrm{dB\ Hz^{2/3}}] &= \frac{2}{3}\left\{\underbrace{10\log\left(I_{\mathrm{rec}}^2 R_b\right)}_{P_{\mathrm{OIP3}}} - \underbrace{\left[10\log\left(I_{\mathrm{rec}}^2 R_b\right) + x_{\mathrm{RIN}} + 10\log\left(\frac{F_A}{4}\right)\right]}_{P_{\mathrm{RIN}}}\right\}\\ &= \frac{2}{3}\left[-x_{\mathrm{RIN}} - 10\log F_A + 10\log 4\right].\end{aligned} \tag{A15}$$

That is the absolute maximum SFDR for a given RIN parameter. For typical parameters, this means SFDR $\leq$ 107 dB Hz$^{2/3}$, which can be seen at high powers in Fig. 8.

In the shot-noise regime, noise increases with optical power, but more slowly that in the RIN regime. SFDR increases with the square root of optical power and gain. The shot-limited SFDR expression is

$$\begin{aligned}\mathrm{SFDR}[\mathrm{dB\ Hz^{2/3}}] &= \frac{2}{3}\left\{\underbrace{10\log\left(I_{\mathrm{rec}}^2 R_b\right)}_{P_{\mathrm{OIP3}}} - \underbrace{\left[10\log\left(I_{\mathrm{rec}} R_b\right) + 10\log\left(\frac{qMF_A}{2}\right)\right]}_{P_{\mathrm{shot}}}\right\}\\ &= \frac{2}{3}\left[10\log\left(I_{\mathrm{rec}}\right) - 10\log\left(\frac{qMF_A}{2}\right)\right],\end{aligned} \tag{A16}$$

where the electron charge, q, evaluates to 191 in log units. Expanding I_{rec}, one finds that the APD gains, M, cancel out.

$$\begin{aligned}&\mathrm{SFDR}[\mathrm{dB\ Hz^{2/3}}]\\ &= \frac{2}{3}\left[10\log P_{\mathrm{pump}} + 10\log\left(\eta_{\mathrm{net}} R_{\mathrm{PD}}\right) - 10\log\left(qF_A\right)\right].\end{aligned} \tag{A17}$$

This means an APD leaves us with only its noise term, F_A, and no benefit in the shot-noise regime.

At low powers, thermal noise dominates. Thermal noise does not scale with I_{rec}, which means more gain (i.e., more received photocurrent) will proportionally increase the SFDR. Here, $P_N \approx P_{\mathrm{thrm}}$, so

$$\mathrm{SFDR}[\mathrm{dB\ Hz^{2/3}}] = \frac{2}{3}\left[\underbrace{10\log\left(I_{\mathrm{rec}}^2 R_b\right)}_{P_{\mathrm{OIP3}}} - \underbrace{10\log\left(k_B T\right)}_{P_{\mathrm{thrm}}}\right]. \tag{A18}$$

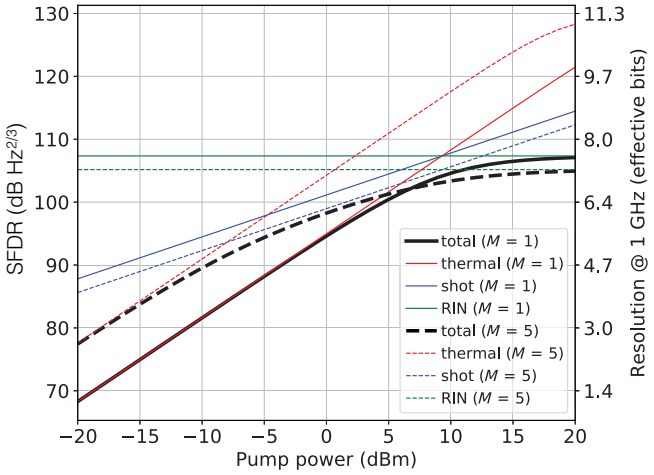

FIG. 8. SFDR components and total for PIN receiver (solid) and APD (dashed). In the thermal regime (red), APD gain can be viewed as a leftwards shift, reducing the input power needed to achieve a given SFDR. An APD strictly degrades shot-limited (blue) and RIN-limited (green) SFDR.

We replace I_{rec} from Eq. (A1) in this low-power, thermal regime in order to state in terms of pump power.

$$\begin{aligned}\text{SFDR[dB Hz}^{2/3}] &= \frac{2}{3}\left[20\log\left(\frac{M\eta_{\text{net}}R_{\text{PD}}P_{\text{pump}}}{2}\right) - 10\log\left(\frac{k_BT}{R_b}\right)\right] \quad \text{(A19)}\\ &= \frac{2}{3}\left[20\log P_{\text{pump}} + 10\log\left(\frac{R_b}{k_BT}\frac{M^2\eta_{\text{net}}R_{\text{PD}}^2}{4}\right)\right]. \quad \text{(A20)}\end{aligned}$$

Using typical values ($R_b = 50\Omega$, $\eta_{\text{net}} = 0.32$, $R_{\text{PD}} = 0.75$ A/W, $T = 290$ K), the second term evaluates to 202 dBHz$^{2/3}$ W^{-2}. Therefore, a pump of 1 mW with a p-i-n diode yields a SFDR of 94.6 dBHz, which is corroborated by the solid red line in Fig. 8.

Analog signal resolution can be stated in terms of an effective number of bits corresponding to an equivalent digital signal, as described around Eq. (9). To give a better practical sense in terms of equivalent bit values: 4 bit, 6 bit, and 8 bit correspond to SFDR of 25.8, 37.9, and 49.9 dB, respectively. At 10 GHz, the frequency contribution subtracts 66.7 dB from the SFDR in dBHz$^{2/3}$. This means, for 4 bits, one would need $25.8 + 66.7 = 92.5$ dBHz$^{2/3}$. For 8 bits, this is $49.9 + 66.7 = 116.6$ dBHz$^{2/3}$, which is not possible, even with an APD, due to the RIN limit in Eq. (A15). The highest frequency that would work at 8 bits within the RIN limit is $10^{3/2\times(107.3-49.9)/10} = 407$ MHz.

APPENDIX B: ELECTRICAL CURRENT IN DETECTOR CIRCUITS

In this Appendix, we derive currents that flow within resistive optical receiver circuits for photonic neural networks, which introduce a key constraint of gain cascadability. In other words, the photodetector has an output swing in terms of voltage. This voltage swing must be enough to drive a voltage-mode modulator at a sufficient strength to change optical power.

Under cascadability constraints, an invariant energy term associated with photodetection will arise. This energy depends strongly on the type of detector: p-i-n versus APD. The implication of results are discussed in Sec. V B. We consider both balanced detection circuits capable of positive and negative weights and single-ended receivers that allow only positive weights. They are pictured in Fig. 9.

Power dissipated in the balanced receiver circuit can be derived from its I-V activity.

$$P_{\text{det}} = \underbrace{i^+(V^+ - V_j)}_{\text{Positive PD}} + \underbrace{i^-(V_j - V^-)}_{\text{Opposing PD}} + \underbrace{(i^+ - i^-)(V_j - V_b)}_{\text{Bias resistor}}, \quad \text{(B1)}$$

where variables pictured in Fig 9(b) are P_{det}, total circuit power; $V^{+/-}$, voltage biases on respective photodiodes; $i^{+/-}$, photocurrent in respective photodiodes; V_j, voltage of the common junction; and V_b, bias voltage on the

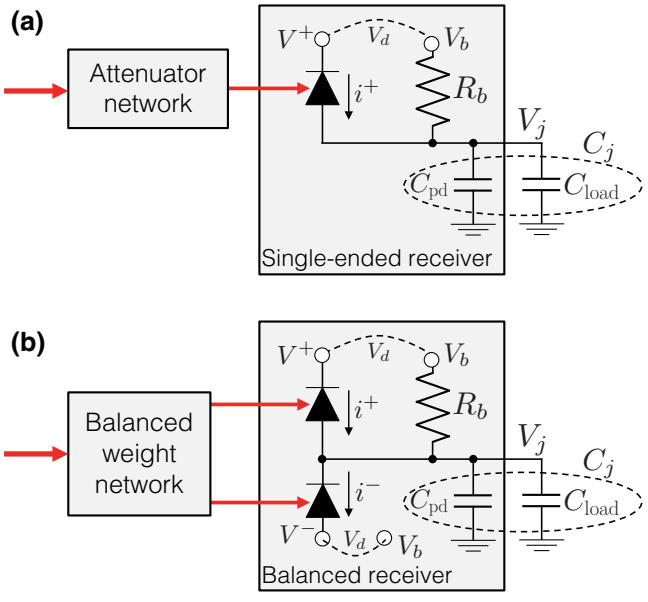

FIG. 9. Photoreceiver circuits showing meanings of variables. Modulators are modeled as a capacitive load in parallel with PD parasitic capacitance. (a) Single-ended circuit capable of positive weighting. (b) Balanced receiver capable of positive and negative (i.e., signal inverting) weighting.

junction. This expression simplifies to

$$P_{\text{det}} = i^{+}(V^{+} - V_b) + i^{-}(V_b - V^{-}). \tag{B2}$$

For convenience, we define a variable, V_d ("dark" voltage), to represent the voltage across a photodetector when no light is present. We also suppose that PD biases are symmetric with respect to junction bias voltage.

$$V_d \equiv (V^{+} - V_b) = (V_b - V^{-}), \tag{B3}$$

$$P_{\text{det}} = (i^{+} + i^{-})V_d, \tag{B4}$$

$$P_{\text{det,single-ended}} = i^{+}V_d, \tag{B5}$$

$$P_{\text{det,balanced}} = i_{\text{max}}V_d. \tag{B6}$$

The difference between the two circuits in Fig. 9 is that single-ended receivers do not have an i^{-}. There is a maximum photocurrent, i_{max}, that depends on the original optical pump power. In a balanced receiver, total photocurrent is always i_{max}. Negative weights are obtained by changing the ratio of light directed to the complementary photodetectors. On average, the power dissipated in the single-ended architecture is half as much. The optical signal is generated by a pump laser, modulated, then passed through a reconfigurable transmission network, which allows us to determine maximum photocurrent

$$i_{\text{max}} = MR_{\text{PD}}\eta_{\text{net}}P_{\text{1pump}}, \tag{B7}$$

where P_{1pump} is the pump power, R_{PD} is the responsivity of each PD, η_{net} is any insertion loss of the network, and M is and photoelectric gain that is present when using avalanche photodetectors.

The cascadability condition defined in Eq. (36) gives a value for pump power. Restating that equation and simplifying i_{max},

$$P_{\text{1pump}}\big|_{g=1} = \frac{2V_\pi}{\pi MR_{\text{PD}}R_b}, \tag{B8}$$

$$i_{\text{max}} = \frac{2V_\pi}{\pi R_b}, \tag{B9}$$

where $g = 1$ is the cascadability condition, and V_π is the modulation slope efficiency of the modulator that is used to generate the optical signal at the input of the transmission network. In an $O/E/O$ neuron, the photoreceiver will also drive this type of modulator.

Bias resistance, R_b, is a free design parameter, but it has an optimal design based on the intended bandwidth of operation.

$$R_b = \left(2\pi C_j f\right)^{-1} \quad \text{(optimal design)}, \tag{B10}$$

$$i_{\text{max}} = 4V_\pi C_j f\,, \tag{B11}$$

where f is operating bandwidth, and C_j is the capacitance of the circuit junction. Junction capacitance is the sum of photodetector parasitic capacitance, C_{pd}, and capacitance of the load, C_{load}. The load can be another modulator, or it could be the input of a digitizing circuit at the back end of an optical subsystem. Substituting into the above equation,

$$E_{\text{det}} \equiv \frac{P_{\text{det}}}{f}, \tag{B12}$$

$$E_{\text{det}} = 4V_\pi C_j V_d, \tag{B13}$$

where we define the variable, E_{det}, to describe the electrical energy associated with detecting modulated optical signals within a cascadable photonic network. Implications of this derivation are discussed in Sec. V B.

[1] E. Strubell, A. Ganesh, and A. McCallum, Energy and policy considerations for deep learning in NLP, ArXiv:1906.02243 (2019).

[2] N. P. Jouppi, *et al.*, In-datacenter performance analysis of a tensor processing unit, ArXiv:1704.04760 (2017).

[3] B. Murmann, D. Bankman, E. Chai, D. Miyashita, and L. Yang, in *2015 49th Asilomar Conference on Signals, Systems and Computers* (IEEE, Pacific Grove, CA, 2015), p. 1341.

[4] S. Jain, A. Sengupta, K. Roy, and A. Raghunathan, RxNN: A framework for evaluating deep neural networks on resistive crossbars, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **40**, 326 (2020).

[5] A. Sengupta, Y. Shim, and K. Roy, Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets, IEEE Trans. Biomed. Circuits Syst. **10**, 1152 (2016).

[6] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[7] P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics* (CRC Press, Boca Raton, FL, 2017).

[8] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Broadcast and weight: An integrated network for scalable photonic spike processing, J. Lightwave Technol. **32**, 4029 (2014).

[9] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, Photonic multiply-accumulate operations for neural networks, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[10] A. R. Totović, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, Femtojoule per MAC neuromorphic photonics: An energy and technology roadmap, IEEE J. Sel. Top. Quantum. Electron. **26**, 1 (2020).

[11] I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, Reprogrammable electro-optic nonlinear activation functions for optical neural networks, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[12] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks, Sci. Rep. **7**, 7430 (2017).

[13] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, Large-Scale Optical Neural Networks Based on Photoelectric Multiplication, Phys. Rev. X **9**, 021032 (2019).

[14] T. Ferreira de Lima, A. N. Tait, H. Saeidi, M. A. Nahmias, H.-T. Peng, S. Abbaslou, B. J. Shastri, and P. R. Prucnal, Noise analysis of photonic modulator neurons, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[15] N. Semenova, X. Porte, L. Andreoli, M. Jacquot, L. Larger, and D. Brunner, Fundamental aspects of noise in analog-hardware neural networks, ArXiv:1907.09002 (2019).

[16] W. Bogaerts, D. Pérez, J. Capmany, D. A. B. Miller, J. Poon, and D. Englund, Programmable photonic circuits, Nature **586**, 207 (2020).

[17] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.17.054029 for a glossary of variables.

[18] V. K. Narayana, S. Sun, A.-H. A. Badawy, V. J. Sorger, and T. El-Ghazawi, MorphoNoC: Exploring the design space of a configurable hybrid NoC using nanophotonics, Microprocess. Microsyst. **50**, 113 (2017).

[19] X. Zheng, E. Chang, P. Amberg, I. Shubin, J. Lexau, F. Liu, H. Thacker, S. S. Djordjevic, S. Lin, Y. Luo, J. Yao, J.-H. Lee, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, A high-speed, tunable silicon photonic ring modulator integrated with ultra-efficient active wavelength control, Opt. Express **22**, 12628 (2014).

[20] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. Shah Hosseini, A. Biberman, and M. R. Watts, An ultralow power athermal silicon modulator, Nat. Commun. **5**, 4008 EP (2014).

[21] Z. Lu, J. Jhoja, J. Klein, X. Wang, A. Liu, J. Flueckiger, J. Pond, and L. Chrostowski, Performance prediction for silicon photonics integrated circuits with layout-dependent correlated manufacturing variability, Opt. Express **25**, 9712 (2017).

[22] L. Chrostowski, X. Wang, J. Flueckiger, Y. Wu, Y. Wang, and S. Talebi Fard, in *Optical Fiber Communication Conference* (OSA, 2014), p. Th2A.37.

[23] M. Soltani, Q. Li, S. Yegnanarayanan, and A. Adibi, Toward ultimate miniaturization of high Q silicon traveling-wave microresonators, Opt. Express **18**, 19541 (2010).

[24] A. Annoni, E. Guglielmi, M. Carminati, G. Ferrari, M. Sampietro, D. A. Miller, A. Melloni, and F. Morichetti, Unscrambling light—automatically undoing strong mixing between modes, Light Sci. Appl. **6**, e17110 (2017).

[25] P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, Thermally tunable silicon racetrack resonators with ultralow tuning power, Opt. Express **18**, 20298 (2010).

[26] J. E. Cunningham, I. Shubin, X. Zheng, T. Pinguet, A. Mekis, Y. Luo, H. Thacker, G. Li, J. Yao, K. Raj, and A. V. Krishnamoorthy, Highly-efficient thermally-tuned resonant optical filters, Opt. Express **18**, 19055 (2010).

[27] J. C. Mikkelsen, W. D. Sacher, and J. K. S. Poon, Adiabatically widened silicon microrings for improved variation tolerance, Opt. Express **22**, 9659 (2014).

[28] P. Alipour, A. H. Atabaki, M. Askari, A. Adibi, and A. Eftekhar, Robust postfabrication trimming of ultracompact resonators on silicon on insulator with relaxed requirements on resolution and alignment, Opt. Lett. **40**, 4476 (2015).

[29] H. Jayatilleka, H. Frish, R. Kumar, J. Heck, C. Ma, M. N. Sakib, D. Huang, and H. Rong, Post-fabrication trimming of silicon photonic ring resonators at wafer-scale, J. Lightwave Technol. **39**, 5083 (2021).

[30] F. Eltes, G. E. Villarreal-Garcia, D. Caimi, H. Siegwart, A. A. Gentile, A. Hart, P. Stark, G. D. Marshall, M. G. Thompson, J. Barreto, J. Fompeyrine, and S. Abel, An integrated cryogenic optical modulator, ArXiv:1904.10902 (2019).

[31] S. Abel, *et al.*, Large pockels effect in micro- and nanostructured barium titanate integrated on silicon, Nat. Mater. **18**, 42 (2019).

[32] $V_\pi L/\Delta V = 4.5/20 = 0.22$ mm.

[33] Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, On-chip photonic synapse, Sci. Adv. **3**, e1700160 (2017).

[34] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities, Nature **569**, 208 (2019).

[35] C. Errando-Herranz, A. Y. Takabayashi, P. Edinger, H. Sattari, K. B. Gylfason, and N. Quack, MEMS for photonic integrated circuits, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[36] C. Errando-Herranz, F. Niklaus, G. Stemme, and K. B. Gylfason, in *2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)* (IEEE, Estoril, Portugal, 2015), p. 53.

[37] P. Edinger, C. Errando-Herranz, and K. B. Gylfason, in *2019 IEEE 32nd International Conference on Micro Electro Mechanical Systems (MEMS)* (2019), p. 919.

[38] R. Baghdadi, M. Gould, S. Gupta, M. Tymchenko, D. Bunandar, C. Ramey, and N. C. Harris, Dual slot-mode NOEM phase shifter, Opt. Express **29**, 19113 (2021).

[39] H. Jayatilleka, K. Murray, M. Ángel Guillén-Torres, M. Caverley, R. Hu, N. AF Jaeger, L. Chrostowski, and S. Shekhar, Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters, Opt. Express **23**, 25084 (2015).

[40] A. Khanna, in *ePIXfab Training Course* (European Conference on Optical Communication, 2015).

[41] D. Marpaung, Ph.D. thesis, University of Twente (2009).

[42] D. Bankman and B. Murmann, Passive charge redistribution digital-to-analogue multiplier, Electron. Lett. **51**, 386 (2015).

[43] F. Akopyan, J. Sawada, A. Cassidy, and R. Alvarez-Icaza, TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **34**, 1537 (2015).

[44] $E_{\text{shot}}(B)$ is always a function of bits, but we sometimes drop the argument for brevity, referring to it as E_{shot}. The same goes for J^*, E_{thrm}, and F_{RIN}.

[45] J. W. Goodman, Fan-in and fan-out with optical interconnections, Opt. Acta: Int. J. Opt. **32**, 1489 (1985).

[46] The s variable describes a concept similar to that described by the ρ variable introduced in Ref. [54] and used in Ref. [9]. In the case referred to as "fixed output precision, only positive inputs and weights" [54, Table 1], there is an embedded, unstated assumption that all signals must be identical, which is always a trivial computation.

[47] X. Xu, M. Tan, B. Corcoran, J. Wu, T. G. Nguyen, A. Boes, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, D. G. Hicks, and D. J. Moss, Single photonic perceptron based on a soliton crystal Kerr microcomb for high-speed, scalable, optical neural networks, ArXiv:2003.01347 (2020).

[48] N. J. D. Martinez, C. T. Derose, R. W. Brock, A. L. Starbuck, A. T. Pomerene, A. L. Lentine, D. C. Trotter, and P. S. Davids, High performance waveguide-coupled Ge-on-Si linear mode avalanche photodiodes, Opt. Express **24**, 19072 (2016).

[49] S. Assefa, F. Xia, and Y. A. Vlasov, Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects, Nat. Lett. **464**, 80 (2010).

[50] P. A. Morton and M. J. Morton, High-power, ultra-low noise hybrid lasers for microwave photonics and optical sensing, J. Lightwave Technol. **36**, 5048 (2018).

[51] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H. B. Miller, H. T. Peng, B. J. Shastri, and P. R. Prucnal, Silicon Photonic Modulator Neuron, Phys. Rev. Appl. **11**, 064043 (2019).

[52] S. Garg, J. Lou, A. Jain, and M. Nahmias, Dynamic precision analog computing for neural networks, ArXiv:2102.06365 (2021).

[53] J. K. George, A. Mehrabian, R. Amin, J. Meng, T. F. de Lima, A. N. Tait, B. J. Shastri, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, Neuromorphic photonics with electro-absorption modulators, Opt. Express **27**, 5181 (2019).

[54] S. Agarwal, T.-T. Quach, O. Parekh, A. H. Hsia, E. P. DeBenedictis, C. D. James, M. J. Marinella, and J. B. Aimone, Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding, Front. Neurosci. **9**, 484 (2016).

[55] A fan-out device where more energy exits than enters has been proposed [13] but is not considered here.

[56] L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. Skirlo, Y. LeCun, M. Tegmark, and M. Soljacic, in *International Conference on Machine Learning*, PMLR, Vol. 70 (2017).

[57] A. N. Tait, A. X. Wu, T. Ferreira de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, Microring weight banks, IEEE J. Sel. Top. Quantum Electron. **22**, 312 (2016).

[58] A. N. Tait, A. X. Wu, T. F. de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Two-pole microring weight banks, Opt. Lett. **43**, 2276 (2018).

[59] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H. T. Peng, P. R. Prucnal, and B. J. Shastri, Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs), IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[60] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, Superconducting Optoelectronic Circuits for Neuromorphic Computing, Phys. Rev. Appl. **7**, 034013 (2017).

[61] E. Timurdogan, C. M. Sorace-Agaskar, E. S. Hosseini, and M. R. Watts, An interior-ridge silicon microring modulator, J. Lightwave Technol. **31**, 3907 (2013).

[62] Z. Ma, M. H. Tahersima, R. Amin, S. Khan, and V. J. Sorger, in *Frontiers in Optics 2017* (Optical Society of America, 2017), p. FM2A.3.

[63] D. Pérez, I. Gasulla, and J. Capmany, Field-programmable photonic arrays, Opt. Express **26**, 27265 (2018).

[64] A. N. Tait, P. Y. Ma, T. F. de Lima, E. C. Blow, M. P. Chang, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Demonstration of multivariate photonics: Blind dimensionality reduction with integrated photonics, J. Lightwave Technol. **37**, 5996 (2019).

[65] M. M. P. Fard, I. A. D. Williamson, M. Edwards, K. Liu, S. Pai, B. Bartlett, M. Minkov, T. W. Hughes, S. Fan, and T.-A. Nguyen, Experimental realization of arbitrary activation functions for optical neural networks, Opt. Express **28**, 12138 (2020).

[66] B. Romeira, J. Javaloyes, C. N. Ironside, J. M. L. Figueiredo, S. Balle, and O. Piro, Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors, Opt. Express **21**, 20931 (2013).

[67] M. A. Nahmias, A. N. Tait, L. Tolias, M. P. Chang, T. Ferreira de Lima, B. J. Shastri, and P. R. Prucnal, An integrated analog O/E/O link for multi-channel laser neurons, Appl. Phys. Lett. **108**, 151106 (2016).

[68] B. Murmann, ADC performance survey 1997–2021, [Online].

[69] A. N. Tait, Quantifying power in silicon photonic neural networks [source code], (2021).

[70] T. Sundstrom, B. Murmann, and C. Svensson, Power dissipation bounds for high-speed Nyquist analog-to-digital converters, IEEE Trans. Circuits Syst. I Regul. Pap. **56**, 509 (2009).

[71] W. Bogaerts and L. Chrostowski, Silicon photonics circuit design: Methods, tools and challenges, Laser. Photon. Rev. **12**, 1700237 (2018).

[72] J. M. Shainline, J. Chiles, S. M. Buckley, A. N. McCaughan, R. P. Mirin, and S. W. Nam, Superconducting optoelectronic neurons V: networks and scaling, ArXiv:1805.01942 (2018).

[73] D. Pérez, I. Gasulla, L. Crudgington, D. J. Thomson, A. Z. Khokhar, K. Li, W. Cao, G. Z. Mashanovich, and J. Capmany, Multipurpose silicon photonics signal processor core, Nat. Commun. **8**, 636 (2017).

[74] G. R. Steinbrecher, J. P. Olson, D. Englund, and J. Carolan, Quantum optical neural networks, ArXiv:1808.10047 (2018).

[75] P. Y. Ma, A. N. Tait, T. F. de Lima, C. Huang, B. J. Shastri, and P. R. Prucnal, Photonic independent component analysis using an on-chip microring weight bank, Opt. Express **28**, 1827 (2020).

[76] Note, signal resolution and weight resolution are different concepts.

[77] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, WRPN: Wide reduced-precision networks, ArXiv:1709.01134 (2017).

[78] J. Schemmel, S. Billaudelle, P. Dauer, and J. Weis, Accelerated analog neuromorphic computing, ArXiv:2003.11996 (2020).

[79] T. Pfeil, T. Potjans, S. Schrader, W. Potjans, J. Schemmel, M. Diesmann, and K. Meier, Is a 4-bit synaptic weight resolution enough?—constraints on enabling spike-timing dependent plasticity in neuromorphic hardware, Front. Neurosci. **6**, 90 (2012).

[80] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, A leaky integrate-and-fire laser neuron for ultrafast cognitive computing, IEEE J. Sel. Top. Quantum Electron. **19**, 1 (2013).

[81] J. M. Shainline, S. M. Buckley, A. N. McCaughan, J. T. Chiles, A. J. Salim, M. Castellanos-Beltran, C. A. Donnelly, M. L. Schneider, R. P. Mirin, and S. W. Nam, Superconducting optoelectronic loop neurons, J. Appl. Phys. **126**, 044902 (2019).

[82] P. R. Prucnal, B. J. Shastri, T. Ferreira de Lima, M. A. Nahmias, and A. N. Tait, Recent progress in semiconductor excitable lasers for photonic spike processing, Adv. Opt. Photonics **8**, 228 (2016).

[83] T. Deng, J. Robertson, and A. Hurtado, Controlled propagation of spiking dynamics in vertical-cavity surface-emitting lasers: Towards neuromorphic photonic networks, IEEE J. Sel. Top. Quantum. Electron. **23**, 1 (2017).

[84] T. V. Vaerenbergh, M. Fiers, P. Mechet, T. Spuesens, R. Kumar, G. Morthier, B. Schrauwen, J. Dambre, and P. Bienstman, Cascadable excitability in microrings, Opt. Express **20**, 20292 (2012).

[85] H. Peng, G. Angelatos, T. F. de Lima, M. A. Nahmias, A. N. Tait, S. Abbaslou, B. J. Shastri, and P. R. Prucnal., Temporal information processing with an integrated laser neuron, IEEE J. Sel. Top. Quantum. Electron. **26**, 1 (2020).

[86] G. Betts, Linearized modulator for suboctave-bandpass optical analog links, IEEE Trans. Microw. Theory Tech. **42**, 2642 (1994).

[87] H. Li, G. Balamurugan, T. Kim, M. N. Sakib, R. Kumar, H. Rong, J. Jaussi, and B. Casper, A 3-D-integrated silicon photonic microring-based 112 Gb/s PAM-4 transmitter with nonlinear equalization and thermal control, IEEE J. Solid-State Circuits **56**, 19 (2021).