Towards Large-Scale Photonic Neural-Network Accelerators

R. Hamerly[1,2], A. Sludds[1], L. Bernstein[1], M. Prabhu[1], C. Roques-Carmes[1],
J. Carolan[1], Y. Yamamoto[2], M. Soljačić[1], and D. Englund[1]
[1]Massachusetts Institute of Technology, Cambridge, MA, email: rhamerly@mit.edu
[2]NTT Research Inc., East Palo Alto, CA

***Abstract*—**Optical approaches to AI acceleration have gained intense interest recently due to the potentially breakthrough advantages of photonics: high bandwidth, low power consumption, and efficient data movement. We overview leading photonic AI platforms based on beamsplitter mesh networks, weight banks, and photoelectric multiplication. While the theoretical performance can be orders of magnitude beyond current state of the art, practical issues of chip area, input / output, and crosstalk paint a more nuanced near-term picture of photonic AI acceleration. Both fundamental and near-term limitations to energy efficiency are addressed, and bandwidth limitations due to temporal crosstalk are analyzed.

I. Introduction

Artificial intelligence (AI) based on deep neural networks (DNNs) has revolutionized many disciplines in computing [1]; however, DNNs are compute- and energy-intensive [2], and limits to available compute are what constrain AI applications in practice. Since DNNs process large amounts of data in regular patterns, special-purpose accelerators have significantly improved the speed and energy consumption compared to CPU or GPU implementations [2-4]. However, challenges with energy consumption [5], the end of Dennard scaling [6], and the looming end of Moore's Law [7] may hinder further performance gains in the long term. This has motivated research into analog or hybrid digital-analog electronic AI accelerators [8-9]. Most accelerators are designed to optimize matrix-matrix multiplication, the bottleneck step [10] in DNN inference (Fig. 1), typically employing memristors and a crossbar array (Fig. 2). Photonics has also emerged as a dark-horse candidate for AI due to its distinct features that offer the possibility of a paradigm shift—high bandwidth limited by optical frequencies, a solution to the interconnect bottleneck [11], and the ability to map linear algebra onto passive matrix multiplication [12]—coupled with recent success at foundry-scale nanophotonic integration [13].

II. Mesh Networks and Weight Banks

The core concepts of analog optical computing [14] and optical neural networks (ONNs) [15] are decades old, but only in recent years has nanophotonics matured to the point that performance competitive with electronics can be contemplated. Like electronic accelerators, ONNs use optics primarily to accelerate the matrix product. Two leading proposals are based on mesh networks [16] and weight banks [17]. In a mesh network, the matrix product is performed by optical interference: signals are encoded in the optical fields entering or leaving the mesh, while the weight matrix is decomposed into a sequence of 2×2 unitary matrices (Fig. 3), implemented with Mach-Zehnder interferometers and tunable phase shifters [12, 18]. In the weight-bank scheme, signals are encoded in the wavelength channels of a single waveguide. This is fanned out to an array of microring resonators that serve as tunable wavelength-dependent splitters ("weight banks"), that separately weight the signal from each wavelength channel (Fig. 4).

Theoretically, the performance of such systems can be quite high. Several factors constrain the system performance in practice. Typically some digital manipulation (e.g. pooling, batch normalization [19]) must be performed on the output data, necessitating A/D and D/A conversion on the inputs and outputs (~1-10 pJ/channel [20]), which indicates large photonic arrays will be required to see significant performance advantages. Tunable photonic devices are quite large [21] ((10-100μm)^2 is typical, see Fig. 5); since an $N\times N$ array requires $O(N^2)$ photonic devices, chip-area constraints will make scaling to the necessary sizes very challenging. Although ONNs have been applied to small, proof-of-principle problems, the goal of a large-scale programmable ONN is as yet unrealized.

III. ONNs Based on Coherent Detection

Recently we proposed an ONN architecture based on coherent detection [22]. The matrix product is decomposed into an array of vector dot products, each of which can be computed using a single homodyne detector (Fig. 6): if two pulse trains encode vectors $\vec{a}$ and $\vec{b}$, the integrated charge will be:

$$Q \propto \int Re[E_a(t)^* E_b(t)]dt \propto \sum_i a_i b_i. \qquad (1)$$

Fig. 7 shows how a matrix-matrix product can be obtained by tiling dot products. For the product $C = AB$, each row of A (resp. column of B) is encoded as a pulse train and fanned out via cylindrical optics to a row (resp. column) of the detector array. This scheme leverages the complementary advantages of free-space optics (spatial multiplexing, fan-out), nanophotonics (large modulator arrays), and dense detector integration [23]. Note that only $O(N)$ photonic modulators are required for each transmitter, significantly alleviating the chip-area constraint and enabling the very large arrays (>10^6 neurons, matrix size $N = 10^3$) needed for next-generation AI workloads.

As with other ONNs, system-level performance of near-term devices will be dominated by I/O costs. For a product of

matrices of dimensions ($m \times k$) and ($k \times n$), the input cost (modulator, DAC) is amortized by the fan-out factor of m (resp. n), while the output cost (detector, ADC) is amortized by the time-integration factor k. The energy per multiply-accumulate (MAC) takes the form $E_{mac} = (m^{-1} + n^{-1})E_{in} + k^{-1}E_{out}$. Table 2 shows estimates of this energy given near-term technology [20, 24], emerging technology [11, 25], and the fundamental Standard Quantum Limit (SQL), which is set by photodetector shot noise (Fig. 8). Fig. 9 shows the theoretical E_{mac} as a function of array size, where a 10^2-10^3 improvement vs. state-of-the-art CMOS is expected with near-term technology. The fact that the SQL dips below the Landauer limit [26] indicates that sub-Landauer performance is in principle possible in ONNs (this is not a contradiction since the Landauer limit only applies to digital, irreversible systems).

IV. Crosstalk

Unwanted crosstalk can degrade the performance of analog optical systems. The mesh-network scheme experiences crosstalk due to imperfect components, which places fairly stringent manufacturing requirements for large systems [27, 28]. In the weight-bank scheme, frequency and neuron count are limited by the time-frequency uncertainty principle $N f_{rep} < B/S$, where B is the optical bandwidth and $S > 1$ is a safety factor.

In the coherent-detection scheme, spatial crosstalk arises because of the close packing of pixels on the detector (Fig. 7). Both diffraction and geometric aberrations contribute to this crosstalk [22], but with appropriate optical engineering near-diffraction-limited focusing is possible. In addition, temporal crosstalk arises if the data rate is close to the modulator bandwidth. Many emerging technologies [29, 30] allow for low-power resonant modulators only with high Q factors and therefore low optical bandwidth (few GHz), so it is desirable to operate as close to the modulator bandwidth as possible. Temporal crosstalk replaces the dot product with a convolution:

$$\sum_i a_i b_i \rightarrow \sum_{i,k} X_{i-k} a_i b_k. \qquad (2)$$

The nearest-neighbor crosstalk X_1 for return-to-zero or non-return-to-zero modulation schemes (Fig. 10) is shown in Fig. 11. MNIST and ImageNet inference are simulated in the presence of crosstalk in Fig. 12. A crosstalk of 5-10% can be tolerated without significant performance degradation, suggesting operation near the modulator's 3-dB cutoff is feasible. Signal pre-emphasis may mitigate crosstalk at higher data rates.

V. Ising Machines

Another compelling application to such hardware is for certain combinatorial optimization problems. Most combinatorial problems belong to the NP-hard complexity class and are thus challenging to solve on conventional processors [31]. Many heuristics based on coupled differential equations, where matrix products are the bottleneck step, show state-of-art performance on such problems [32-34]. Recently we proposed and demonstrated a proof-of-concept optical "Ising machine" based on parametric oscillator networks [35], but with electrically mediated spin-spin couplings (Fig. 13). On many benchmark problems, the system outperforms sparsely connected quantum annealers such as D-Wave (Fig. 14) [36]. Significant performance gains may be possible if optical couplings based on ONN hardware are utilized (Fig. 15) [37, 38].

VI. Conclusion

Photonics offers a new path to solve the compute problem in deep learning. While component density will never rival electronics, photonic systems benefit from high bandwidth, low-loss propagation through dielectrics, and potentially very low energy consumption. Approaches based on mesh networks and weight banks show promise but suffer from chip-area limitations. We have introduced an approach based on coherent detection that solves the chip-area problem and promises significant energy-efficiency benefits over the current state-of-art. Crosstalk simulations show high data rates limited by modulator speeds are possible. Beyond deep learning, such accelerators find use in NP-hard optimization problems.

Acknowledgments

This research is supported by: IC Postdoctoral Fellowship at MIT (DOE / ODNI), NSERC Doctoral Fellowship, NSF GRFP, U.S. ARO at ISN / MIT (no. W911NF-18-2-0048), and in-kind support form NVIDIA. The authors acknowledge Vivienne Sze and Joel Emer for helpful discussions.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, Nature, vol. 521, p. 436, 2015.
[2] V. Sze, Y.-H. Chen *et al.*, Proc. IEEE, vol. 105, no. 12, p. 2295, 2017.
[3] N. P. Jouppi, C. Young, N. Patil *et al.*, 2017 ACM/IEEE ISCA, pp. 1-12.
[4] B. Zimmer, R. Venkatesan *et al.*, 2019 Symp. VLSI Tech., p. C24-1.
[5] M. Horowitz, 2014 IEEE ISSCC, pp. 10-14.
[6] R. H. Dennard *et al.*, IEEE J. Solid-State Circ., vol. 9, no. 5, p. 256, 1974.
[7] G. E. Moore, Electronics, pp. 114–117, 1965.
[8] S. Ambrogio *et al.*, Nature, vol. 558, no. 7708, pp. 60–67, 2018.
[9] S. George et al., IEEE Trans. VLSI Systems, vol. 24, no. 6, p. 2253, 2016.
[10] B. Fleischer *et al.*, 2018 Symp. VLSI Circuits, p. C4-2.
[11] D. A. B. Miller, J. Lightw. Technol., vol. 35, no. 3, pp. 346-396, 2017.
[12] M. Reck *et al.*, Phys. Rev. Lett., vol. 73, no. 1, p. 58, 1994.
[13] E. Timurdogan, Z. Su, C. Poulton, M. Byrd *et al.*, 2018 OFC, p. M3F.1.
[14] A. Vander Lugt, IEEE Trans. Info. Theory, vol. 10, pp. 139-145, 1987.
[15] E. Paek and D. Psaltis, Opt. Engineering, vol. 26, no. 265428, 1987.
[16] Y. Shen, N. C. Harris *et al.*, Nature Photonics, vol. 11, no. 7, p. 441, 2017.
[17] A. N. Tait, T. F. Lima *et al.*, Sci. Rep., vol. 7, no. 1, p. 7430, 2017.
[18] W. R. Clements *et al.*, Optica, vol. 3, no. 12, pp. 1460-1465, 2016.
[19] I. Goodfellow *et al.*, *Deep learning.* MIT Press, 2016
[20] B. Jonsson, 2010 IEEE ICECS, pp. 766-769.
[21] N. C. Harris *et al.*, Opt. Express, vol. 22, no. 9, pp. 10487-10493, 2014.
[22] R. Hamerly et al., Phys. Rev. X, vol. 9, no. 2, p. 021032, 2019.
[23] A. Rogalski, Prog. Quant. Electron., vol. 36, no. 2, pp. 342-473, 2012.
[24] DARPA, BAA No. HR001119S0004, Nov. 2018.
[25] M. Notomi, K. Nozaki *et al.*, Opt. Comm., vol. 314, pp. 3-17, 2014.
[26] R. Landauer, IBM J. Res. Dev., vol. 5, no. 3, pp. 183-191, 1961.
[27] R. Burgwal *et al.*, Opt. Express, vol. 25, no. 23, p. 28236, 2017.
[28] M. Y.-S. Fang *et al.*, Opt. Express, vol. 27, no. 10, p. 140009, 2019.
[29] C. Wang, M. Zhang *et al.*, Nature, vol. 562, no. 7725, p. 101, 2018.
[30] C. Koos *et al.*, J. Lightw. Technol., vol. 34, no. 2, pp. 256-268, 2016.
[31] S. Rudich *et al.*, *Computational Complexity Theory*. AMS, 2014.
[32] T. Leleu *et al.*, Phys. Rev. Lett., vol. 122, no. 4, p. 040607, 2019.
[33] K. Kalinin and N. Berloff, Sci. Rep., vol. 8, no. 1, p. 17791, 2018.
[34] B. Molnar and M. Ercsey-Ravasz, PLOS One, vol. 8, p. e73400, 2013.
[35] P. L. McMahon *et al.*, Science, vol. 354, no. 6312, pp. 614-617, 2016.
[36] R. Hamerly, T. Inagaki *et al.*, Sci. Adv., vol. 5, no. 5, p. eaau0823, 2019.
[37] C. Roques-Carmes, Y. Shen, C. Zanoci *et al.*, arXiv:1811.02705, 2018.
[38] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris *et al.*, *in preparation*.

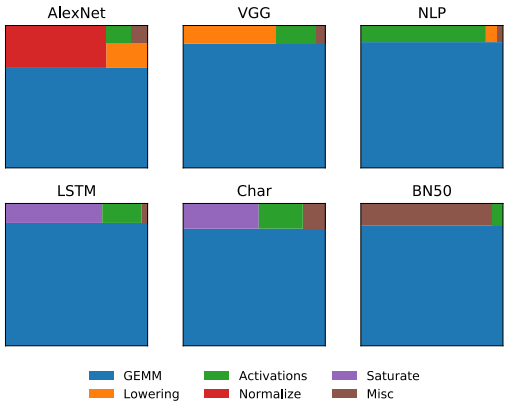

Fig. 1. Breakdown of computational costs in typical deep learning workloads. Matrix-matrix products (GEMM) typically account for 80-90% of the total [10].

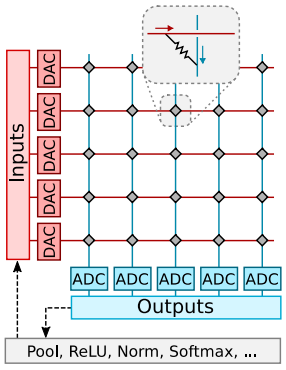

Fig. 2. Crossbar architecture for resistive memory-based analog matrix-vector multiplication.

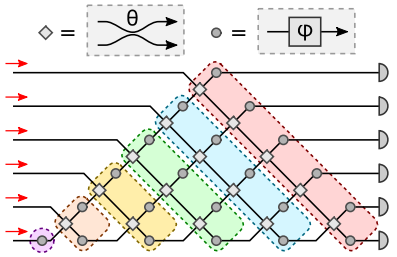

Fig. 3. Mesh-network ONN implementation of matrix-vector product [12, 16].

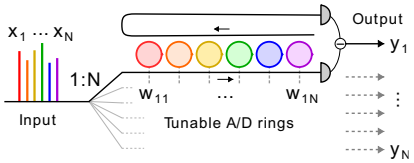

Fig. 4. Weight-bank ONN scheme [17].

Type	Concept	Company	Throughput (TMAC/s/cm^2)	Energy (pJ/MAC)	Stage	Challenges
Digital	GPU	Nvidia, AMD	8	10	Commer-	Addressed. Very
	ASIC	Multiple	14	1	cial	mature technology.
Analog electronics	Memristors	Multiple	> 1000	< 0.001	Proto-	Updates, noise, etc.
	FPAA	–	> 100	0.01	type	Chip area, variations
Photonics	Mesh	Lightmatter	> 100	0.001–0.1	Proto-	Area, matrix size
	Weight bank	Luminous	> 100	0.001–0.1	type	# channels, size
This work	Homodyne	n/a	> 1000	0.001	Concept	Unknown

Table. 1. Comparison of optical and electronic neural-network accelerator approaches.

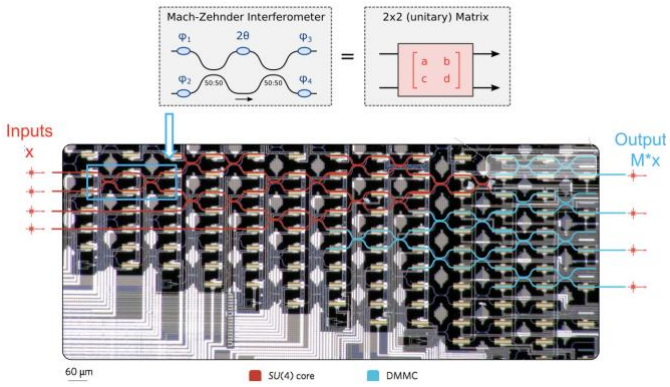

Fig. 5. Image of mesh network fabricated to represent 4×4 programmable matrix-vector product [16].

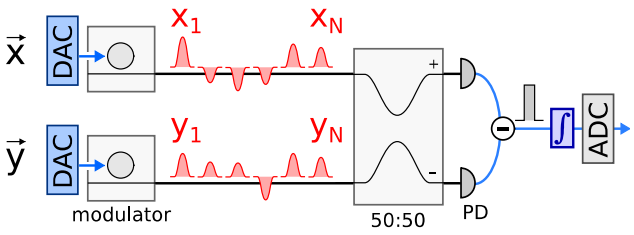

Fig. 6. Vector dot product by coherent detection. Vectors encoded on optical pulse trains using modulators. A 50:50 beamsplitter mixes the signals. The integrated charge gives the product: $Q \propto \sum_n x_n y_n$.

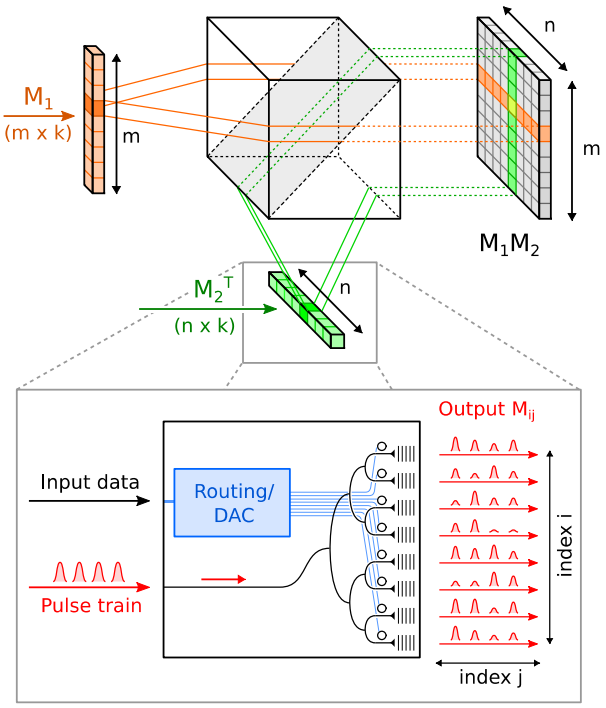

Fig. 7. Schematic of optical matrix-product accelerator based on coherent detection [22]. Integrated modulator arrays convert data to optical pulse trains. Cylindrical lenses (not shown) provide fan-out to rows / columns of detector array, which computes product.

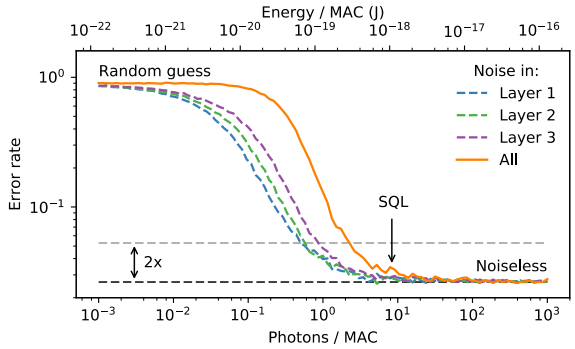

Fig. 8. Effect of quantum noise on error rate of MLP100 neural network (MNIST, two N=100 inner layers) as a function of optical energy per MAC, showing the standard quantum limit (SQL) [22].

Problem	NN	Energy E_{mac}		
		pJ I/O	fJ I/O	SQL
MNIST	MLP-100	13 fJ	13 aJ	1 aJ
	MLP-1000	3 fJ	3 aJ	0.1 aJ
ImageNet	AlexNet	8 fJ	8 aJ	3 aJ

Table 2. ONN energy efficiency estimates for three benchmark problems. Figures based on near-term technology (picojoule-scale modulators, detectors, ADC) [24] and far-term technology (femtojoule-scale modulators / detectors) [25] are plotted against the SQL.

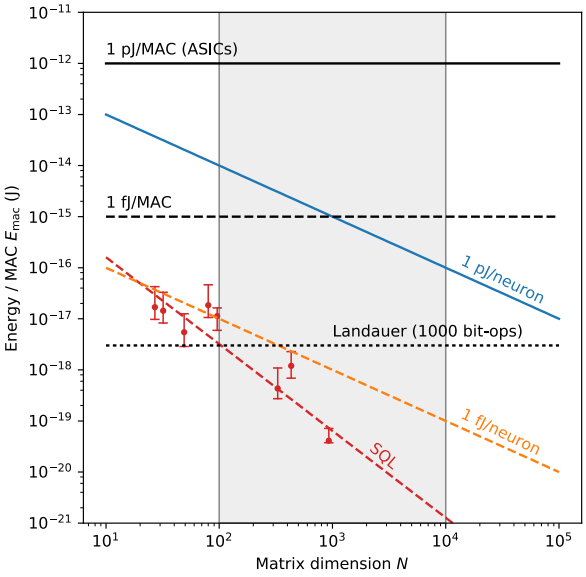

Fig. 9. Plot of limits to the energy consumption for coherent detection based ONN: O(pJ/neuron) bound for near-term I/O technology, O(fJ/neuron) due to emerging technology, and quantum limit [22]. Most problems have matrix sizes $10^2 < N < 10^4$ (shaded region).

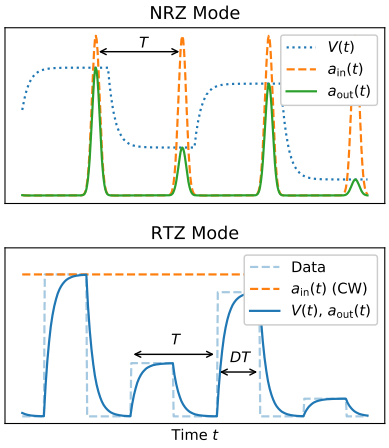

Fig. 10. NRZ and RTZ modes of modulator operation to reduce temporal crosstalk. RC-limited case is shown.

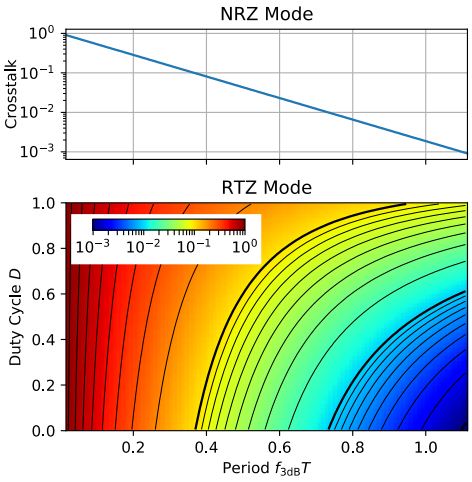

Fig. 11. Temporal crosstalk as a function of pulse spacing, normalized to modulator 3dB bandwidth.

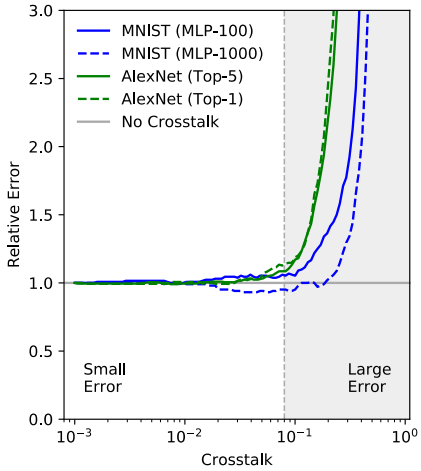

Fig. 12. ONN error in presence of temporal crosstalk (normalized to crosstalk-free case).

Fig. 13. Annealing machines for Ising combinatorial optimization. Left: CIM-based LASOLV from NTT. Right: D-Wave Systems 2000Q based on quantum annealing.

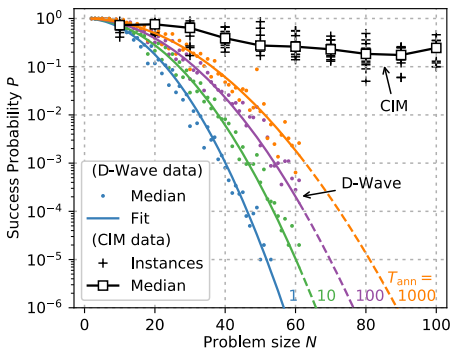

Fig. 14. CIM and D-Wave 2000Q success probabilities at SK benchmark problems [36].

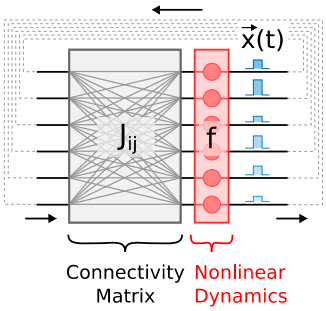

Fig. 15. Principle of photonic recurrent Ising sampler exploiting ONN matrix-product accelerator [37-38].