Invited: Neuromorphic architectures based on augmented silicon photonics platforms

Matěj Hejda
Hewlett Packard Labs, Hewlett Packard Enterprise
Diegem, Belgium
matej.hejda@hpe.com

Federico Marchesin
Ghent University - imec
Ghent, Belgium
federico.marchesin@ugent.be

George Papadimitriou
Department of Informatics and Telecomm., University of Athens
Athens, Greece
georgepap@di.uoa.gr

Dimitris Gizopoulos
Department of Informatics and Telecomm., University of Athens
Athens, Greece
dgizop@di.uoa.gr

Benoit Charbonnier
Univ. Grenoble Alpes, CEA, LETI
Grenoble, France
benoit.charbonnier@cea.fr

Régis Orobtchouk
Univ. Lyon, ECL, INSA-Lyon, UCBL, CPE Lyon, CNRS, INL
Lyon, France
regis.orobtchouk@insa-lyon.fr

Peter Bienstman
Ghent University - imec
Ghent, Belgium
peter.bienstman@ugent.be

Thomas Van Vaerenbergh
Hewlett Packard Labs, Hewlett Packard Enterprise
Diegem, Belgium
thomas.van-vaerenbergh@hpe.com

Fabio Pavanello
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, INPG, CROMA
Grenoble, France
fabio.pavanello@cnrs.fr

ABSTRACT

In this work, we discuss our vision for neuromorphic accelerators based on integrated photonics within the framework of the Horizon Europe NEUROPULS project. Augmented integrated photonic architectures that leverage phase-change and III-V materials for optical computing will be presented. A CMOS-compatible platform will be discussed that integrates these materials to fabricate photonic neuromorphic architectures, along with a gem5-based simulation platform to model accelerator operation once it is interfaced with a RISC-V processor. This simulation platform enables accurate system-level accelerator modeling and benchmarking in terms of key metrics such as speed, energy consumption, and footprint.

CCS CONCEPTS

• **Hardware → Neural systems**; • **Applied computing → Physics**; • **Computing methodologies → Modeling and simulation**.

KEYWORDS

Artificial Intelligence, Photonics, Computing, Neuromorphic, AI accelerators, gem5, Accelerators modeling, Simulation

ACM Reference Format:
Matěj Hejda, Federico Marchesin, George Papadimitriou, Dimitris Gizopoulos, Benoit Charbonnier, Régis Orobtchouk, Peter Bienstman, Thomas Van Vaerenbergh, and Fabio Pavanello. 2024. Invited: Neuromorphic architectures based on augmented silicon photonics platforms. In *Proceedings of 61st ACM/IEEE Design Automation Conference (DAC '24)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advances in deep learning (and more recently genAI) empower machines with remarkable information processing and synthesis capabilities. At the same time, the exponential growth in data volumes generated by an extensive range of consumer devices and industrial sensors calls for novel approaches to advanced and efficient data processing. Although advanced AI models are typically deployed in purpose-built high-performance cloud clusters, some applications require advanced data processing locally at the edge, that is, closer to where data were originally generated [3]. More specifically, edge computing requires the development of lightweight accelerators capable of providing AI-tailored data processing locally with low latency and high energy efficiency [19]. Current state-of-the-art AI is enabled by graphical processing units (GPUs), tensor processing cores and similar application-specific hardware that allows for high computing parallelization and optimized computation of linear algebraic operations that underpin modern deep learning workloads [18]. In line with other conventional computing approaches, these digital architectures utilize Von Neumann processor architectures with distinct memory and computing blocks. This poses some inherent limitations in terms of high demands on data movement between these functional blocks in terms of bandwidth, energy efficiency, and power consumption.

The goal of alleviating these bottlenecks is one of the main objectives in fields such as in-memory computing, where (a part of) the data needed for the computation is co-located with the computing core architecture, thus requiring less data movement [23]. In neuromorphic (brain-inspired) engineering, the principle

DAC '24, June 23–27, 2024, San Francisco, CA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

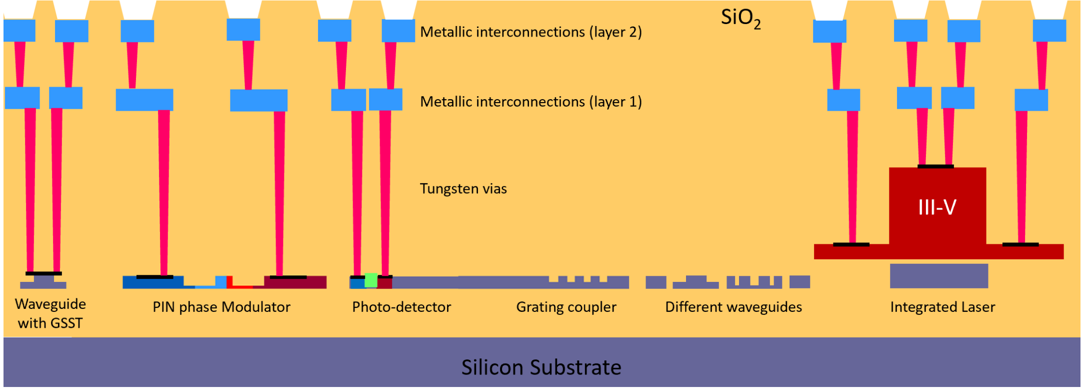

Figure 1: CMOS-compatible augmented platform in the NEUROPULS project. Reprinted with permissions from [15].

of memory-compute co-location is further enhanced with signaling and information processing principles analogous to those observed in the brain, which promises further improvements in computation sparsity and energy efficiency [11]. Although significant efforts have been devoted to leveraging existing CMOS technology to develop digital and mixed-signal neuromorphic computing cores [6, 13], alternative technologies such as spintronics or photonics are nowadays gaining considerable momentum for the implementation of next-generation computing hardware [12, 20]. However, there are still certain practical limitations towards achieving accelerators based on integrated photonics. Current CMOS-compatible silicon photonic platforms do not provide a full range of required functional modalities. In addition, photonic accelerators are often investigated standalone, rather than being interfaced with a system-level architecture and used in real-world scenarios. Finally, full-scale, system-level simulation platforms of a photonic accelerator interfaced with a processor core remain mostly unexplored. In the next sections, we will discuss how we are aiming to tackle all these aspects within the framework of the NEUROPULS project.

2 CMOS-COMPATIBLE SIPH PLATFORM

Unlike in the electronic neuromorphic approaches, photonics allow one to leverage desirable properties of lightwaves such as wavelength multiplexing, low-loss signal propagation without Joule heating as well as access to very high bandwidth devices (above tens of GHz) to e.g. encode or read-out data [22]. In particular, integrated photonics has emerged as a promising size, weight and power (SWaP)-optimized platform. Silicon photonics (SiPh) currently represents arguably the most promising approach to photonic integration due to its compatibility with existing CMOS approaches for cost and volume. However, certain functionalities are not available in pure Silicon-On-Insulator (SOI) platforms. In particular, the ability to realize non-volatile optical memory elements is one of such missing functionalities that is key to achieving energy-efficient optical computing architectures [8, 9]. In addition, active devices (such as lasers) cannot be realized in silicon because of its indirect bandgap. Therefore, additional materials such as III-V compound semiconductors need to be co-integrated into SOI platforms, typically using heterogeneous or hybrid integration methods. A monolithic fabrication approach is the most desirable way to achieve these missing functionalities in a compact manner with excellent alignment tolerances between patterned layers and consequent ease of coupling, as well as packaging costs lower than those of hybrid approaches. Such platform has not yet been developed or presented as an open access service, as is the case for pure SOI platforms [16].

One of the goals of the Horizon Europe NEUROPULS project is to develop a platform that can accommodate these additional functionalities in a monolithic manner. In Figure 1, the integration strategies for PCMs (e.g., GSST or other types) and III-V materials are shown. This integration does not affect other building blocks, such as high-speed modulators and detectors (above 50 GHz bandwidth) that have already been developed for the platform and are already provided to the end users. Therefore, novel building blocks taking advantage of these additional functionalities will be developed to further enhance the existing selection of building blocks on the SOI platform.

3 PCM/III-V AUGMENTED SOI BUILDING BLOCKS

One of the key building blocks for broadband neuromorphic photonic architectures is the Mach-Zenhder interferometer (MZI, shown in Figure 2(a)). An individual MZI consists of couplers and phase-shifters (PS) and represents a $SU(2)$ transformation. Typically in SOI, a specific phase-shift is induced through the thermo-optic effect via an adjacent heater, and continuously consumes electrical power. Given that this phase-shift remains constant for a set weight matrix (that is, during inference), a non-volatile approach would be ideal to remove this constant energy consumption [14]. Furthermore, besides the non-volatile nature of optical phase shift or attenuation effects, the devices should be compact with minimized optical loss to enable deep arrangements of MZIs. One of the goals of NEUROPULS is the development of low-loss, compact, and reconfigurable multilevel PCM-based MZIs with heaters above PCM patches and waveguides (see Figure 2(a)). Various approaches are currently being investigated to benefit from the properties of PCMs such as GeSe and GSST that present a large figure of merit (FOM) given by $\delta n/\delta k$, where δn and δk are the refractive index contrasts for the real and imaginary part, respectively, around the standard telecom wavelength of 1550 nm [7, 21].

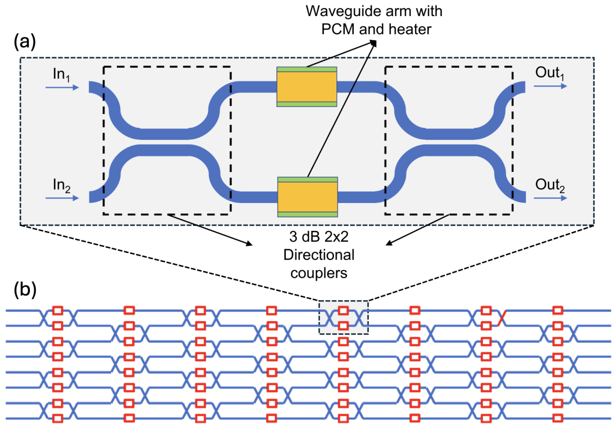
Waveguide arm with PCM and heater
(a)
In_1
In_2
Out_1
Out_2
3 dB 2x2 Directional couplers
(b)

Figure 2: (a) MZI with the PCM-augmented (in green) non-volatile optical phase-shifters with heaters on top (in yellow) for programmability. (b) An example of an MZI mesh architecture (here implementing 8×8 matrix) dedicated to accelerating matrix-vector multiplication operations via in-memory optical computing.

Furthermore, Q-switched III-V on-chip lasers are explored as chipscale excitable spiking sources. Pioneering works in this direction have already been carried out; however, spikes were generated off-chip, unlike the approach that we will be pursuing in NEUROPULS [9]. By leveraging the ultrafast response (sub-ns) and accumulation behavior of PCM-based devices to optical pulses, the viability of photonic spiking neural networks (SNN) and bio-inspired learning rules such as spike-timing dependent plasticity (STDP) will be investigated.

4 COMPUTING ARCHITECTURES

The main focus of the photonic neuromorphic architecture is to realize optical in-memory acceleration of linear algebra operations that underpin a majority of current deep learning models. In particular, we focus on realizing a photonic matrix-vector multiplication (MVM) engine to enable a generalized matrix-matrix (GeMM) core. These architectures are based on meshes of programmable integrated MZIs that operate as multiport interferometers with a degree of matrix expressivity (universality) determined by component arrangement. Within the NEUROPULS project, various mesh architectures of MZIs (or, more generally, couplers and phase-shifters) are investigated and evaluated. These include previously proposed mesh architectures such as the Clements [5] architecture with compacted interferometers [1] (shown in Figure 2(b)) or the Fldzhyan [10] architecture with parallel PS blocks [1], as well as newly proposed multiport interferometer architectures. In these, input vectors are encoded into amplitude/phase of individual inputs (typically using high-speed Mach Zehnder modulators), and the multiplication (weighting) matrix is encoded in the state of the programmable PS blocks. Generalization to GeMM operations can be realized through separating of the input matrix into rows, and processing those either via time-division multiplexing or through encoding into multiple dense wavelength division multiplexed (DWDM) channels that can be processed in parallel in a single multiport interferometer without incurring additional resource costs.

5 SIMULATION PLATFORM

In the ever-evolving landscape of computing systems, the integration of neuromorphic accelerators and hardware security primitives on a consolidated simulation platform is crucial. Both electronic and photonic domain-specific accelerators (DSAs) continue to be active areas of research and development as the demand for specialized and efficient computing solutions grows across various industries

In the NEUROPULS project, we will: (a) create efficient full system simulation tools on top of the gem5 simulator [2] to model and evaluate complete computing systems with neuromorphic accelerators and security primitives, as shown in Figure 3; (b) explore the diverse design space of heterogeneous computing systems employing photonic neuromorphic accelerators and hardware primitives; and (c) facilitate detailed system-level evaluation of both software and hardware, with a specific emphasis on the security properties of the computing platform.

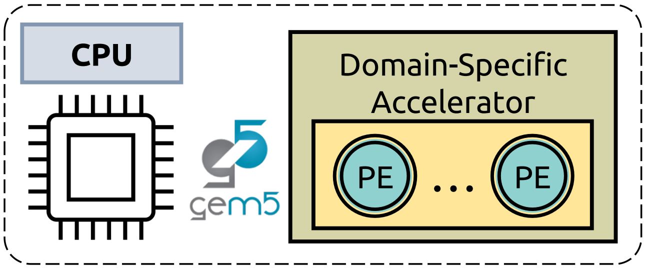

Figure 3: gem5-based system architecture modeling and simulation infrastructure overview.

NEUROPULS simulator is based on gem5, an open-source, system-level computer architecture simulator that provides a flexible and modular framework for modeling and simulating various aspects of computer systems. It supports cycle-level simulation of a wide range of computer architectures, including x86, Arm, RISC-V, and others, making it a versatile platform at the system level (including the microarchitecture, architecture, operating systems, and application layers of the computing stack). Given the tremendous growth of the RISC-V ecosystem in the past few years, we have taken gem5-SALAM, which uses an advanced dynamic graph execution engine based on LLVM [17] and only supports Arm ISA processor cores (the CPU part in Figure 3), and ported it to support the RISC-V ISA and system configuration. We introduce gem5-MARVEL [4], which is based on LLVM IR (Intermediate Representation) to model DSAs using C descriptions of their functionality. The gem5-based simulation infrastructure comprises two core components: the Compute Unit and the Communications Interface. The Compute Unit represents the custom accelerator's datapath, while the Communications Interface facilitates memory access, control, and synchronization through memory access ports, Memory-Mapped Registers (MMRs), and interrupt lines. The memory access ports allow parallel access to different memory types, such as scratchpad memories (SPMs) and register banks (these two types of memories occupy the largest part of the area of many accelerators). MMRs consist of configurable status, control, and data registers, allowing low-level device configuration and facilitating communication between the accelerator and the host, as well as between multiple accelerators (i.e., processing elements - PEs) in a cluster (as shown in the right-most side of Figure 3). By treating the accelerator as a memory-mapped device, the host can utilize the provided interrupt signals for synchronization without the need for constant polling. Additionally, the gem5-based infrastructure includes Direct Memory Access (DMA) devices and custom memories that can be seamlessly integrated into accelerator designs, enhancing its versatility. gem5-MARVEL is also a fault injection framework that operates at the microarchitecture level and supports transient and permanent fault injections to all hardware structures of the CPU and for the three prevailing ISAs (Arm, x86, RISC-V). The fault injection feature was implemented in the simulation framework to support the reliability aspect of the NEUROPULS project.

6 CONCLUSIONS

Integrated photonics represent one of the technological platforms with potential to empower modern heterogeneous computing systems by enabling high bandwidth and improved energy efficiency during data movement and computing. While silicon photonics benefits from compatibility with CMOS technology, further augmentation with additional material platforms is required to unlock the full scale of needed chipscale optical building blocks. In the NEUROPULS project, we have augmented a CMOS-compatible SOI photonic platform with III-V semiconductor technology (enabling on-chip active optical devices), and chalcogenide-based PCMs (to realize non-volatile optical modulators). Using this augmented platform, we propose and evaluate a system-level implementation of a neuromorphic accelerator based on an in-memory photonic MVM accelerator core. Various MZI mesh architectures are evaluated for the MVM core, including their performance, matrix expressivity and robustness. Furthermore, we have developed a novel gem5-based simulation framework with RISC-V ISA support to allow for extensive performance evaluation and benchmarking of the complete photonic-enabled accelerator interfaced with controllers and processors. We believe this comprehensive system-level implementation is key to realize practical, photonic neuromorphic accelerator.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101070238. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] B. A. Bell and I. A. Walmsley. 2021. Further Compactifying Linear Optical Unitaries. *APL Photonics* 6, 7 (July 2021), 070804. https://doi.org/10.1063/5.0053421

[2] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, et al. 2011. The Gem5 Simulator. *SIGARCH Comput. Archit. News* 39, 2 (aug 2011), 1–7. https://doi.org/10.1145/2024716.2024718

[3] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. 2020. An Overview on Edge Computing Research. *IEEE Access* 8 (2020), 85714–85728. https://doi.org/ACCESS.2020.2991734 Conference Name: IEEE Access.

[4] Odysseas Chatzopoulos, George Papadimitriou, Vasileios Karakostas, and Dimitris Gizopoulos. 2024. gem5-MARVEL: Microarchitecture-Level Resilience Analysis of Heterogeneous SoC Architectures. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA 2024)*. 543–559. https://doi.org/10.1109/HPCA57654.2024.00047

[5] William R. Clements, Peter C. Humphreys, Benjamin J. Metcalf, W. Steven Kolthammer, and Ian A. Walsmley. 2016. Optimal design for universal multiport interferometers. *Optica* 3, 12 (Dec. 2016), 1460. https://doi.org/10.1364/OPTICA.3.001460

[6] Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A. Fonseca Guerra, et al. 2021. Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook. *Proc. IEEE* 109, 5 (May 2021), 911–934. https://doi.org/10.1109/JPROC.2021.3067593 Conference Name: Proceedings of the IEEE.

[7] J.-B. Dory, C. Castro-Chavarria, A. Verdy, J.-B. Jager, M. Bernard, et al. 2020. Ge–Sb–S–Se–Te amorphous chalcogenide thin films towards on-chip nonlinear photonic devices. *Sci Rep* 10, 1 (July 2020), 11894. https://doi.org/10.1038/s41598-020-67377-9

[8] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, et al. 2021. Parallel convolutional processing using an integrated photonic tensor core. *Nature* 589, 7840 (Jan. 2021), 52–58. https://doi.org/10.1038/s41586-020-03070-1

[9] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice. 2019. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* 569, 7755 (May 2019), 208–214. https://doi.org/10.1038/s41586-019-1157-8

[10] S. A. Fldzhyan, M. Yu Saygin, and S. P. Kulik. 2020. Optimal Design of Error-Tolerant Reprogrammable Multiport Interferometers. *Optics Letters* 45, 9 (May 2020), 2632–2635. https://doi.org/10.1364/OL.385433

[11] Charlotte Frenkel, David Bol, and Giacomo Indiveri. 2023. Bottom-Up and Top-Down Approaches for the Design of Neuromorphic Processing Systems: Tradeoffs and Synergies Between Natural and Artificial Intelligence. *Proc. IEEE* 111, 6 (June 2023), 623–652. https://doi.org/10.1109/JPROC.2023.3273520

[12] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, et al. 2020. Neuromorphic spintronics. *Nat Electron* 3, 7 (July 2020), 360–370. https://doi.org/10.1038/s41928-019-0360-9 Publisher: Nature Publishing Group.

[13] Sebastian Höppner, Yexin Yan, Andreas Dixius, Stefan Scholze, Johannes Partzsch, et al. 2022. The SpiNNaker 2 Processing Element Architecture for Hybrid Digital Neuromorphic Computing. http://arxiv.org/abs/2103.08392

[14] Mario Miscuglio, Jiawei Meng, Omer Yesiliurt, Yifei Zhang, Ludmila J Prokopeva, et al. [n. d.]. Artificial Synapse with Mnemonic Functionality using GSST-based Photonic Integrated Memory. ([n. d.]), 8.

[15] Fabio Pavanello, Cedric Marchand, Ian O'Connor, Regis Orobtchouk, Fabien Mandorlo, et al. 2023. NEUROPULS: NEURomorphic energy-efficient secure accelerators based on Phase change materials aUgmented siLicon photonicS. *2023 IEEE European Test Symposium (ETS)* (May 2023). https://doi.org/10.1109/ETS56758.2023.10173974

[16] Abdul Rahim, Thijs Spuesens, Roel Baets, and Wim Bogaerts. 2018. Open-Access Silicon Photonics: Current Status and Emerging Initiatives. *Proc. IEEE* 106, 12 (Dec. 2018), 2313–2330. https://doi.org/10.1109/JPROC.2018.2878686 Conference Name: Proceedings of the IEEE.

[17] Samuel Rogers, Joshua Slycord, Mohammadreza Baharani, and Hamed Tabkhi. 2020. gem5-SALAM: A System Architecture for LLVM-based Accelerator Modeling. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 471–482. https://doi.org/10.1109/MICRO50266.2020.00047

[18] Wojciech Samek, Gregoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Muller. 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* 109, 3 (March 2021), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

[19] Mahadev Satyanarayanan. 2017. The Emergence of Edge Computing. *Computer* 50, 1 (Jan. 2017), 30–39. https://doi.org/10.1109/MC.2017.9

[20] Bhavin J. Shastri, Alexander N. Tait, Thomas Ferreira de Lima, Wolfram H. P. Pernice, Harish Bhaskaran, et al. 2021. Photonics for Artificial Intelligence and Neuromorphic Computing. *Nature Photonics* 15, 2 (Feb. 2021), 102–114. https://doi.org/10.1038/s41566-020-00754-y

[21] Richard Soref, Joshua Hendrickson, Haibo Liang, Arka Majumdar, Jianwei Mu, et al. 2015. Electro-optical switching at 1550 nm using a two-state GeSe phase-change layer. *Opt. Express* 23, 2 (Jan. 2015), 1536. https://doi.org/10.1364/OE.23.001536

[22] Alexander N. Tait, Thomas Ferreira de Lima, Mitchell A. Nahmias, Heidi B. Miller, Hsuan-Tung Peng, et al. 2019. Silicon Photonic Modulator Neuron. *Phys. Rev. Appl.* 11, 6 (June 2019), 064043. https://doi.org/10.1103/PhysRevApplied.11.064043 Publisher: American Physical Society.

[23] Wen Zhou, Bowei Dong, Nikolaos Farmakidis, Xuan Li, Nathan Youngblood, et al. 2023. In-memory photonic dot-product engine with electrically programmable weight banks. *Nat Commun* 14, 1 (May 2023), 2887. https://doi.org/10.1038/s41467-023-38473-x