NEUROPULS: NEUROmorphic energy-efficient secure accelerators based on Phase change materials aUgmented siLicon photonicS
NEUROPULS: NEUROmorphic energy-efficient secure accelerators based on Phase change materials aUgmented siLicon photonics

Fabio Pavanello1, Cedric Marchand1, Ian O’Connor1, Régis Orobjouchk4, Fabien Mandorlo1, Xavier Letartre1, Sebastien Cueff1, Elena Ioana Vatajelu2, Giorgio Di Natale2, Benoit Cluzel3, Aurelien Coillet3, Benoit Charbonnier4, Pierre Noe4, Frantisek Kavan5, Martin Zoldak5, Michal Szaj5, Peter Bientman6, Thomas Van Vaerenbergh7, Ulrich Ruhmair8, Paulo Flores9, Luis Guerra e Silva9, Ricardo Chaves9, Luis-Miguel Silveira9, Mariano Ceccato10, Dimitris Gizopoulos11, George Papadimitriou11, Vasileios Karakostas11, Axel Brando12, Francisco J. Cazorla12, Ramon Canal12,13, Pau Closas14, Adrià Gusi-Amigó14, Paolo Crovetti15, Alessio Carpegna16, Tzann Melendez Carmona16, Stefano Di Carlo16, Alessandro Savino16

1Univ. Lyon, Ecole Centrale de Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CPE Lyon, CNRS, INL
2Univ. Grenoble Alpes, CNRS, Grenoble INP, TIMA, 38000 Grenoble, France
3ICB UMR CNRS 6303, Université de Bourgogne Franche-Comté, Dijon, France
4Univ. Grenoble Alpes, CEA, LETI, Grenoble, France, 5ARGOTECH, Náchod, Czech Republic
6Ghent University - imec, Gent, Belgium, 7Hewlett Packard Labs, HPE Belgium, B-1831 Diegem, Belgium
8Physics Dept. LMU Munich, Munchen, Germany, 9INESC-ID Lisboa, Lisbon, Portugal
10Department of Computer Science, University of Verona – Verona, Italy
11Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
12Barcelona Supercomputing Center, Barcelona, Spain, 13Universitat Politècnica de Catalunya, Barcelona, Spain
14Albora Technologies SL, Barcelona, Spain
15Politecnico di Torino, Dept. of Electronics and Telecommunications (DET), Italy
16Politecnico di Torino, Control and Computer Eng. Department, Italy

Abstract—This special session paper introduces the Horizon Europe NEUROPULS project, which targets the development of secure and energy-efficient RISC-V interfaced neuromorphic accelerators using augmented silicon photonics technology. Our approach aims to develop an augmented silicon photonics platform, an FPGA-powered RISC-V-connected computing platform, and a complete simulation platform to demonstrate the neuromorphic accelerator capabilities. In particular, their main advantages and limitations will be addressed concerning the underpinning technology for each platform. Then, we will discuss three targeted use cases for edge-computing applications: Global National Satellite System (GNSS) anti-jamming, autonomous driving, and anomaly detection in edge devices. Finally, we will address the reliability and security aspects of the stand-alone accelerator implementation and the project use cases.

Index Terms—artificial neural networks, modeling, simulation

I. INTRODUCTION

The increasing need to process large amounts of data has been a major driver for developing novel, energy-efficient computing architectures [1]. Among the multitude of approaches to tackle energy efficiency requirements, brain-inspired neuromorphic architectures are one of the most viable solutions thanks to their inner absence of I/O bottleneck between memory and processing units and their more natural mapping to machine learning (ML) algorithms [2]. Neuro-morphic approaches are especially suitable for developing lightweight, low-latency, and low-power accelerators in line with the requirements sought by edge-computing systems for applications such as autonomous driving, Internet of Things (IoT) devices, and 5G networks.

Integrating photonics can efficiently target all the needs above [3] among the various technologies for building neuro-morphic architectures. This is mainly due to the current availability of mature CMOS-compatible platforms, e.g., Silicon-on-Insulator (SOI), which allow the dense integration of devices such as modulators, detectors, optical interfaces, etc., for volume scaling. In particular, photonic approaches can provide a key advantage concerning electronic ones regarding energy consumption, bandwidth, and latency thanks to low propagation losses in photonic integrated circuits (PICs) as well as the possibility for large parallelism using wavelength-multiplexing techniques, high-speed data encoding/decoding devices, and speed-of-light propagation throughout photonic
In what follows, we will provide an overview of the NEUROPULS project and its scientific objectives towards strengthening the European digital supply chain and developing European secure specialized accelerator designs capable of delivering high-performance computing at ultra-low power operation. NEUROPULS’ unique technology will improve the performance per watt by at least two orders of magnitude for the targeted edge applications, thus addressing the performance reduction of current CMOS-based general-purpose computing platforms and the slow-down of Moore’s law.

II. PROPOSED APPROACH IN NEUROPULS

In NEUROPULS, we will develop three different platforms at technological, hardware computing, and simulation levels (see Fig. 1). The technological platform is based on the open-access silicon photonics platform from CEA-LETI with the addition of phase-change materials (PCMs) and III-V materials. The building blocks based on these technologies, such as plastic non-volatile photonic synapses [5] and non-linear neurons based on spiking lasers [6], will enable various kinds of neuromorphic photonic architectures that NEUROPULS will investigate.

Security layers will be built upon Physical Unclonable Functions (PUFs), security primitives that avoid digital keys stored in memory [7]. These primitives will be embedded in the photonic and electronic chips. A software/hardware interface will leverage the security layers at a silicon level to set up multiple security features for the accelerator, such as secure authentication, encryption of neural network (NN) data, and signing of the output data.

Using high-speed RF wiring and packaging, the photonic architectures will be driven by an ASIC chip linked to the PIC chip. The developed accelerator will then be interfaced by an FPGA where a RISC-V processor is implemented. The processor will be the interface between the outside world and the accelerator. It would allow transmission/retrieval of raw data/computing results and will handle the security protocols, leveraging the security primitives. In our vision, the ASIC chip will also contain the RISC-V processor for large-volume manufacturing, leading to a stand-alone self-consistent hardware computing platform. However, for prototyping reasons, the ASIC in NEUROPULS will only contain the driving circuitry, the interface with the FPGA, and electronic PUFs. The latter will be used for hardware integrity to bundle the photonic and electronic chips uniquely. The PIC and the ASIC chips will be mounted on a board and connected using an FPGA interface to the FPGA hosting a RISC-V processor.

Computational models with different complexity levels for PCM-based photonic devices and systems will also be developed to simulate the behavior of the neuromorphic and PUF architectures taped out in the silicon photonics platforms and their response once included within the entire computing platform. Finally, we will use the accelerator for three use cases of interest in edge computing: GNSS anti-jamming, autonomous driving, and anomaly detection in edge devices. State-of-the-art simulation tools considering the whole hardware computing platform will allow us to investigate performance scaling for the accelerator, e.g., energy consumption, latency, and speed as a function of the number of inputs/outputs/layers/neurons, etc. well its security features. In NEUROPULS, we target highly demanding performance from our accelerator, which will allow us to improve by, e.g., two orders of magnitude the energy/MAC metric compared to concurring GPU technology for the selected edge-computing use cases. Besides, security features based on PUFs will be investigated for the stand-alone accelerator and the targeted use cases using both hardware and software approaches, from the security primitives level to the software-driven security protocols implemented.

III. CMOS-COMPATIBLE PLATFORM

The platform that will be developed in the project is based on the silicon photonics platform from CEA-LETI, available in open-access multi-project wafer runs. It already includes several key building blocks, such as optical interfaces (grating couplers), modulators, detectors, resonators, etc. (see Fig. 2).
However, one of the novel features we aim to integrate includes PCMs. These thin PCM layers, e.g., GeSbTe (GST) and GeSSbTe (GSST) materials with different stoichiometric compositions, will be placed as thin (10-20-nm-thick) rectangular patches above optical waveguides as shown in Fig. 2. Their integration influences the optical field propagation due to changes in effective refractive index depending on the PCM state, i.e., amorphous versus crystalline. Strong modulation of the optical properties can be obtained due to the optical field interaction with the PCM patch. Various waveguides will be available, namely strip (fully etched) and rib (partially etched) configurations, where the overall silicon layer thickness is 300 nm. Two layers of metal interconnect will also be present to ease the routing of the components, and heaters will also be included (not present in the cross-section) to modify the crystalline degree of the PCM. Thanks to high-performance traveling-wave modulators and Ge detectors, the platform allows operating at frequencies above 50 GHz.

IV. HARDWARE COMPUTING PLATFORM

Regarding the optical architecture, we will implement different options based on combining silicon photonics and PCMs.

One straightforward option is to consider an accelerator for matrix-vector products in feedforward NNs (FFNNs), where a mesh of modulators is programmed so that the system implements a certain matrix multiplication [4]. This can be done, e.g., by using singular value decomposition to factor the matrix into the product of a unitary matrix, a diagonal matrix, and another unitary matrix.

The main limitation here is that the optical components’ size prevents the implementation of huge matrices. Therefore, we will also explore other options for implementing larger systems. One approach we will consider is pruning, which involves removing unnecessary connections and weights in the network to reduce the overall size of the matrix. Another option is to use block matrix decompositions, where the matrix is divided into smaller blocks that can be processed separately. This allows for larger matrices to be implemented with a smaller number of optical components.

Furthermore, we will explore tensor-train approaches proposed in [8]. These approaches involve representing the matrix as a product of low-rank tensors, which can be more efficiently processed using optical components. Using these various approaches, we aim to develop a scalable optical architecture that can handle larger matrices and more complex NNs. This will enable faster, more scalable, and more efficient training of NNs using photonics. Apart from feedforward architectures, we will also study recurrent NNs (RNNs). We can consider fully trainable RNNs, but a variant called reservoir computing will also be considered. In reservoir computing, an RNN is randomly initialized and left untrained. Instead, what is trained is a linear combination of the time traces of the signals at each node. This technique has the advantage of being more computationally efficient than fully trainable RNNs, while still achieving good performance in many applications. This approach has been implemented in integrated photonics before [9]. It has been used, e.g., to realize nonlinear dispersion compensation of telecom signals [10], demonstrating the potential of photonics in a wide range of applications beyond traditional ML.

We plan to incorporate PCMs in our proposed architecture to implement non-volatile optical weights. Various studies have used these materials, such as [11] and [12]. Using non-volatile weights significantly reduces power consumption compared to volatile weights, which must be driven continuously or refreshed periodically. However, in addition to their non-volatility, PCMs have another advantage we plan to exploit in our architecture: their nonlinear dynamics. E.g., by exciting the material with pulses rather than continuous-wave excitation, we can take advantage of the nonlinear behavior of the material to enable other computing paradigms, such as spiking NNs (SNNs). In such NNs, the neurons communicate using brief pulses or spikes rather than continuously varying signals. This enables the implementation of energy-efficient and highly parallel NNs. To generate the spikes injected into the system, we will monolithically integrate lasers in III-V materials in the same platform to develop advanced high-extinction ratio (ER > 8 dB) Q-switched spiking lasers [13]. These hybrid III-V-on-Si spiking lasers are a scalable and low-cost alternative to
previous pure III-V versions of these Q-switched lasers.

V. SIMULATION PLATFORM

Novel hardware architectures require the development of an ecosystem of state-of-the-art tools capable of supporting and promoting innovative hardware approaches. Specifically, the feasibility of innovation demands simulation tools that allow exploring the functional benefits of the new technology while supporting a precise estimation of the impact of the accelerators on the final system regarding performance, power consumption, and reliability, among others.

The NEUROPULS project aims to develop a simulation platform that models and evaluates computing systems incorporating neuromorphic accelerators and hardware security primitives. The simulation platform will explore the design space of heterogeneous computing systems that utilize photonic neuromorphic accelerators and hardware primitives to achieve optimal performance. The project will design and implement a toolchain for modeling and simulation at the system level to make photonics accelerators functional and programmable. The toolchain will automate the process of generating system-level models of photonic modules with varying levels of complexity and accuracy. The simulation platform will model a complete computing system, including multiple CPU cores, memory hierarchy, PNNs, and photonic security primitives (PUFs-based). The platform will also cover all computing stack layers, from hardware to application software. To evaluate the security provisions of photonics technology, the simulation platform will consider all software security aspects at the system and application software layers.

The platform will be built around gem5 [14], [15]: a state-of-the-art microarchitecture-level simulator widely used in many studies [14–16]. It will be enhanced with recent extensions supporting RISC-V-based systems modeling with a rich set of accelerators and flexible interfaces. The platform’s flexibility will ease the exploration of trade-offs between security levels and corresponding performance, power consumption, and reliability penalties.

NEUROPULS will also investigate the reliability and power estimation of photonic hardware components. The NEUROPULS simulation platform aims to provide a flexible infrastructure that enables the comparison of various photonic accelerator parameters to determine neuromorphic hardware’s power reduction and performance improvements. In addition, the platform will focus on RISC-V CPU cores and customized neuromorphic accelerators, along with their corresponding software stack, to evaluate the reliability of the complete heterogeneous system and identify potential reliability issues that different NN accelerators may face. Therefore, NEUROPULS will create a simulation platform that is rapid, adaptable, and modular. It will be constructed using the gem5 simulator as its foundation and will support the evaluation of advanced NN accelerators and PUFs.

VI. TARGETED USE-CASES

A. GNSS anti-jamming

Satellite-based navigation systems, called GNSS, provide an accurate and reliable solution to position, velocity, and time (PNT) estimation in a myriad of applications [17]. However, GNSS can be compromised by L-band intentional/unintentional jamming interferences, which typically overpower a GNSS receiver intending to deny its operation [18]. Jamming episodes are not rare nowadays, making jammers cheap to buy and/or easy to build [19], which threatens safety-critical applications or critical infrastructures relying on GNSS PNT. Therefore, interference management is crucial to resilient GNSS usage, including detecting and classifying jamming occurrences. In this context, it has been shown in [20], [21] that ML models provide promising results, including setups where those models are learned in a distributed federated learning manner [22]. NEUROPULS targets this use case by enabling efficient NN-based classifiers to warn users of the presence/type of jamming interference.

B. Autonomous driving

Trajectory prediction is fundamental to autonomous driving frameworks such as Apollo [23] and Autoware [24]. Some trajectory prediction systems build to different extents on RNNs. For instance, Apollo uses some such networks as part of its trajectory prediction system. Other implementations building on Long Short-Term Memory (LSTM) are realized using multi-input multi-output LSTM-based RNNs [25]. Therefore, efficient RNN implementations bring new opportunities for trajectory prediction in autonomous driving systems. NEUROPULS will use its acceleration technology to explore the efficiency of RNN-based trajectory prediction systems. The aim is to start with the simpler RNNs part of Apollo, mainly to test the approach’s viability, and then scale to more complex LSTM-based RNNs to test the limits of the approach in the context of trajectory prediction.

C. Anomaly detection in edge devices

As an example of in-network computing using the photonic neuromorphic accelerator developed in the NEUROPULS project, we will target security applications by investigating the potential of NNs emulated by the photonic hardware to perform anomaly detection in a resource-constrained edge environment. Having these ultrafast neuromorphic accelerators embedded at the Edge to flag anomalies would enable identifying the malicious events where more expensive post-processing would otherwise be required. Possible use cases could be the prevention of IoT botnet, DDoS attacks [26], [27], or in-vehicle network security [28]. By nature, anomaly detection is event-driven. Consequently, it is potentially a good match with the temporal information processing enabled by SNNs [28], [29], one of the NN architectures investigated during the NEUROPULS project. The aim is first to predict the performance of a variety of promising anomaly detection algorithms when adapted to the constraints present in photonic neuromorphic hardware using the hardware models.
and simulation framework developed in the project, where we will target benchmark tasks that will allow us to compare simulated classification accuracy, predicted energy-efficiency and problem size scalability with the reported performance of existing state-of-the-art algorithms running on traditional digital hardware as well as alternative emerging neuromorphic accelerators [30].

VII. RELIABILITY AND SECURITY ASPECTS

In NEUROPULS, we will develop security layers at the silicon level based on PUFs. More precisely, we will leverage the already available technology for the photonic accelerator to implement novel architectures for building CMOS-compatible photonic PUFs. Such implementations are expected to be more robust against ambient fluctuations, aging, side-channel attacks, and machine-learning modeling than PUFs based on electronic technologies. This is due to the unique properties of PICs, such as lower dependence on temperature fluctuations or EM interference. Most of their components, such as waveguides, do not present aging issues nor signal leakages throughout propagation (light is confined in a tight area below dielectric layers), and a large number of degrees of freedom (phase, polarization, amplitude, mode number, etc.) is available to enhance system complexity [31].

Our research will investigate both so-called Weak PUFs and Strong PUFs [32]: The Weak PUF approach will use photonic PUF designs with very few challenges per PUF. A system-specific secret key or a system-specific (secret key, public key)-pair will be derived from the (noisy) responses of these Weak PUFs. Key aspects here are the already mentioned long-term stability of the responses, the use of optimal techniques for key extraction and error correction, and the statistical independence of the PUF responses for neighboring or adjacent photonic Weak PUFs. To verify the latter, statistical tools like the NIST suite will be employed [33]. Another important aspect is the secure processing of the derived key within the accelerator and the surrounding electronic system and circuits. In the Strong PUF approach, a more complex photonic PUF structure must and will be used. As required by Strong PUFs, it shall have a very large number of challenge-response pairs (CRPs) and a highly complex challenge-response (or input-output) relation. Non-linear optical effects inside the photonic PUF must be exploited in this context.

We will systematically test their response for complexity to verify the sufficient complexity (and thus security) of our photonic Strong PUF designs. Several existing Strong PUF tests will be applied, and various new tests will be developed simultaneously. They assess aspects such as the higher-order non-linearities in the PUF-responses, challenge bit sensitivity, or pseudo-randomness as measured by the NIST suite [33]. Another key strategy for evaluating the security and complexity of our photonic Strong PUFs is the application of ML algorithms [34]. Given a relatively small set of PUF-CRPs, we will assess whether ML algorithms can be trained to predict other yet unknown CRPs. Critical figures here are the number of CRPs used in the ML algorithm’s training phase, the computational effort in the training phase, and the prediction accuracy of the trained ML model. We will apply different ML strategies, including the latest NNs and related techniques. We stress that on the protocol and application side, the two primitives of Weak PUFs and Strong PUFs shall be mixed and interleaved, combining their mutual strengths: Strong PUFs will allow identification and authentication of messages with long-term or short-term digital keys in the system. Weak PUFs can be employed to derive secret digital keys whenever needed, at least avoiding the long-term presence of digital secrets in the hardware [35].

Such a high level of security at the hardware level will be exploited to secure the software level, i.e., the application code that the RISC-V processor runs. This will be achieved by designing well-defined APIs at the hardware–software interface to let the software layer request security services from the hardware layer. These services include encryption, decryption, and digital signing tasks, based on secret cryptographic keys that never leak to the software and thus to a potential malicious analyst [36]. On top of this, strong software hardening solutions would be developed, such as program obfuscation [37] and tamper-detection [38] to limit malicious reverse engineering attacks to the software components.

VIII. CONCLUSIONS

The NEUROPULS project will address several research areas across the supply chain, from material science and photonic technologies to computing architectures, security layers, and high-level simulations. Three different platforms will be developed: a silicon photonics-powered chip, an FPGA-powered RISC-V-connected platform, and a complete simulation ecosystem to demonstrate the neuromorphic accelerator capabilities. The development of such platforms will be pivotal for the NEUROPULS project to showcase a low-power secure accelerator based on an augmented silicon photonics platform addressing three different use cases of interest for edge computing: GNSS anti-jamming, autonomous driving, and anomaly detection in edge devices. We will target two orders of magnitude lower power consumption concerning state-of-the-art technology for the selected use cases. Novel security layers based on the developed CMOS-compatible photonic platform will be developed by tightly integrating security primitives, i.e., PUFs, and security software interfaces. Such integration will enable various security features, which we will apply to the accelerator as stand-alone and for specific use cases. The modular approach we propose in NEUROPULS will inform us about the technology scaling potential without relying solely on hardware prototypes but by accurate system modeling using gem5-based simulators. The project’s outcomes will foster an augmented interest in this technology and spur novel research directions and investments at a European level in photonics, which has been indicated as a key enabling technology for Industry 4.0 and RISC-V processors technology.
REFERENCES


