PHOTONICS PUBLIC PRIVATE PARTNERSHIP

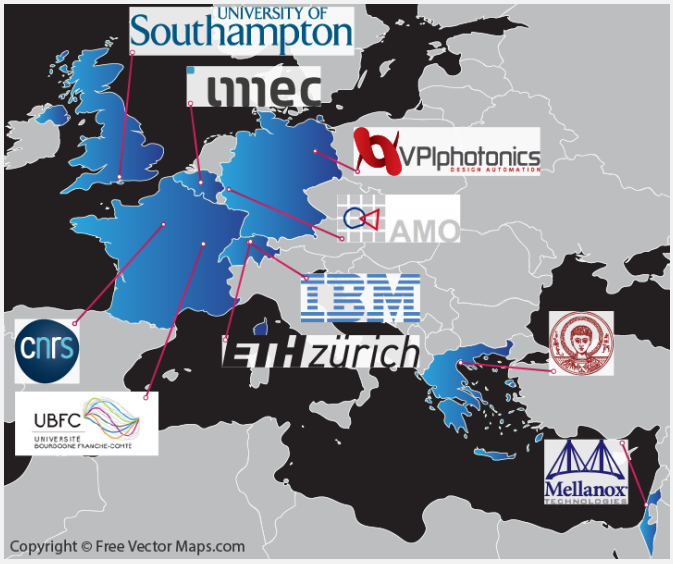
plasmoniAC

Energy- and Size-Efficient Ultra-Fast Plasmonic Circuits for Neuromorphic Computing Architectures

Grant Agreement	871391
Programme	Horizon 2020 –Work Programme 2018-2020 – Information and Communication Technologies: Unconventional Nanoelectronics
Duration	01/01/2020 – 31/12/2022 (36 Months)
Budget	Overall Cost: € 4,114,926.25 EU Contribution: € 3,999,458.75
Coordinator	Aristotle University of Thessaloniki, GR

Contact	**Prof. Nikolaos Pleros,** email: npleros@csd.auth.gr **Dr. George Dabos,** email: ntamposg@csd.auth.gr **Dr. Angelina Totovic,** email: angelina@auth.gr

Website	http://www.plasmoniac.eu



The challenge

Computing industry is rapidly moving from a programming to a learning era, where the 10 GMAC/s/W (MAC – multiply-accumulate operation) digital energy efficiency wall of current CMOS electronics and von-Neumann architectures simply cannot keep the pace with the new computational power metrics required. The breakdown of Moore's and Koomey's laws referring to CPU and energy efficiency scaling is only validating the industrial consensus: by 2020, the reign of the von-Neumann architecture will begin fading away after 75 years of dominance, with non-von-Neumann layouts expected to be the key-enablers for optimizing Deep Learning (DL) and particularly inferencing, reducing the energy requirements for many computational classes. This new computing paradigm has already begun to unfold, leading to the development of large neuromorphic machines that already exceed the energy and size-efficiency walls of classical platforms.

However, the size and energy advantages of electronic processors are naturally counteracted by the speed and power limits of the electronic interconnects in the circuits due to RC parasitic effects, preventing further improvement in energy efficiency and computational power required for unleashing the huge potential of Deep Neural Networks and Artificial Intelligence (AI). This reality has already triggered a new research field – neuromorphic photonics – which aims to transfer the well-known high-bandwidth and low-energy interconnect credentials of photonic circuitry in the area of neuromorphic platforms. However, prevailing the interconnect industry does certainly not imply that photonics is the suitable technology when computations are required: having the critical dimensions of few mm's and/or consuming > 25 mW even for weighing functionality.

Being currently at the dawn of neuromorphic computing, a future-proof solution that could dominate this landscape for many years should obviously rely on the best-performing and top-efficient technology mixture that can support the complete DL learning model portfolio. This is exactly where plasmonics comes to bridge what electronics and photonics lack: the size of electronics and the speed of photonics.

Mission statement

Plasmonics has been proposed as the key technology that may exploit its metallic nature to mix optical functions with low-dimension electronics, offering a natural platform for synergizing photonic-level bandwidths with electronic-level sizes within an ultra-high energy efficiency envelope. **PlasmoniAC** aims to take advantage of these fundamental benefits towards deploying and demonstrating a neuromorphic plasmonic platform that can optimize speed, energy and size efficiency across all its constituent circuitry by utilizing and advancing the best-in-class technology and material platforms. **PlasmoniAC** will release a whole new class of energy- and size-efficient feed-forward (FF) and recurrent artificial plasmonic neurons with up to 100 GHz

clock frequencies and 1 and 6 orders of magnitude better energy- and footprint-efficiencies, smoothly synergizing:

- Si_xN_y-based photonics as its Nx100 Gb/s interconnect layer for low-loss transmission up to mm-long distances,
- ultra-low energy and μm^2-scale plasmonic weights as its computational layer with record-high $>10^{14}$ MAC/s/W energy efficiency per element, and
- nm-size memristor electronics for use as weight control and reconfigurability circuitry.

Project objectives

PlasmoniAC invests in neuromorphic computing towards sustaining processing power and energy efficiency scaling, adopting the best-in-class material and technology platforms for optimizing computational power, size and energy at every of its constituent functions. Following a holistic hardware/software co-design approach, **PlasmoniAC** targets the following objectives:

i) to elevate plasmonics into a computationally-credible platform with Nx100 Gb/s bandwidth, μm^2-scale size and $>10^{14}$ MAC/s/W computational energy efficiency, using CMOS compatible BTO and SiOC materials for electro- and thermo-optic computational functions,
ii) to blend them via a powerful 3D co-integration platform with Si_xN_y-based photonic interconnects and with non-volatile memristor-based weight control,
iii) to fabricate two different sets of 100 Gb/s 16- and 8-fan-in linear plasmonic neurons,
iv) to deploy a whole new class of plasmo-electronic and nanophotonic activation modules,
v) to demonstrate a full-set of $\sin^2(x)$, ReLU, sigmoid and tanh plasmonic neurons for FF and recurrent neurons,
vi) to embrace them into a properly adapted Deep Learning training model suite, ultimately delivering a neuromorphic plasmonic software design library, and
vii) to apply them on IT security-oriented applications for threat and malware detection.

Target technology breakthroughs

Plasmonic modulators: The backbone of the **PlasmoniAC** computational circuitry will be a set of novel modulators atop of low-loss Si_xN_y waveguide platform: (*i*) strongly E/O BTO operating at 100 Gb/s with $V_\pi L$ as low as 150 µmV and losses of 0.5 dB/µm, and (*ii*) strongly T/O SiOC with power consumption <8 mW and total losses <6 dB for π phase shift.

Technology platform synergy: Harnessing the advantages of best-in-class platforms, **PlasmoniAC** aims to employ Si_xN_y photonic waveguides as a low-loss and high-bandwidth interconnect platform, relying on non-volatile memristor nanotechnology to enable neuron programmability and reconfigurability, finally bridging the two by introducing plasmonic layer for compact and low-energy neuromorphic computational functions. Within this goal, PlasmoniAC will develop a 3D cointegration process, smoothly extending from bulky photonic structures via µm-scale plasmonic elements up to nanometer size memristor electronics, designating a layer for each discrete technology platform.

Programmable linear plasmonic neurons: PlasmoniAC aims to deliver two alternative reconfigurable linear plasmonic neuron architectures, operating at 100 Gb/s: (*i*) a 16-fan-in wavelength-encoded linear plasmonic neuron promising an energy- and size-efficiency of $\sim 5\times10^4$ GMAC/s/W (0.02 pJ/MAC) and $> 10^{10}$ MMAC/s/cm^2 that are 1 and >6 orders of magnitude higher, respectively, compared to the state-of-the-art neuromorphic machines, and (*ii*) an 8-fan-in phase-encoded linear plasmonic neuron targeting the performance of 0.8×10^{12} MAC/s per neuron, yielding a computational- and a size-efficiency of 2×10^4 GMAC/s/W (0.05 pJ/MAC) and $\sim 4\times10^9$ MMAC/s/cm^2, close to 1 and 5 orders of magnitude better than the state-of-the-art electronic machines, respectively.

Activation modules: Recognizing the need for deploying a variety of ultra-fast nonlinear activation functions at up to 100 Gb/s, **PlasmoniAC** will make a leap forward in combining the speed and functionality by demonstrating: (*i*) a 100 Gb/s plasmo-electronic module capable of providing ReLU and rectified $\sin^2(x)$ activation functions targeting -6 dBm sensitivity, 400 mW electrical power consumption and 0.8 mm^2 footprint; (*ii*) a nanophotonic sigmoid and tanh(x) activation function implemented by an InP-on-Si_xN_y PhC injection locked laser operating at 50 Gb/s, occupying only 75 μm^2 and consuming <0.5 mW; (*iii*) a self-feedback sigmoid activation function that enables the realization of recurrent neuron layouts with a footprint of 0.2×0.2 μm^2 and <1 mW power consumption.

Feed-forward and recurrent plasmonic neurons: Combining its linear neurons with a broad set of custom nonlinear activation modules, **PlasmoniAC** will validate its neuromorphic technology platform through (*i*) WDM-accelerated plasmo-electronic neuron in 100 Gb/s $\sin^2(x)$ and ReLU layouts and (*ii*) phase-encoded FF sigmoid and tanh, as well as recurrent plasmo-photonic neurons.

Deep Learning models and algorithms: PlasmoniAC will follow the path from technology development to addressing real application needs by developing a new set of DL training models and algorithms, incorporating noise, cross-talk and hardware nonidealities, concluding to optimal initialization and weight regularization schemes.

Data Center IT security: Real-time classification of infected data packets within Data Centers at up to 100 GHz will be a ground for showcasing **PlasmoniAC**'s hardware/software co-design approach, striving towards co-location of its neuromorphic hardware on top of the future Data Center routing hardware in an attempt to catch the threat far from the servers and prevent malicious access to user data.

Software design library: In securing the take-up of its new neuromorphic hardware platform, **PlasmoniAC** will embed its new technology into ready-to-use libraries within VPIphotonics Design Suite™ environment.