2

Technology and Hardware for Neuromorphic Computing

Björn Debaillie, Ilja Ocket, and Peter Debacker

imec, Belgium

Abstract

Edge artificial intelligence and machine-learning algorithms increasingly enter our day-to-day products and applications. This massive adoption of data in all aspects of human activity will lead to unprecedented growth in computational needs to process this data into useful information and actions. The current approach to process this data in high-end cloud server parks is no longer sustainable as it costs energy, latency, and poses privacy threats. Realizing intelligent energy-efficient local processing is however extremely challenging. Neuromorphic computing, modelled according to the human's brain nerve network, is often suggested to realize such processing. Building such neuromorphic processing hardware however requires major advancements at different levels. New technology platforms for emerging semiconductor devices must be developed, levering emerging memory technologies which show characteristics related to neuromorphic computation. MRAM (Magnetoresistive Random Access Memory) could mimic the stochastic behavior of synapses, FeRAM (Ferroelectric Random Access Memory) could be tuned to emulate synaptic weight, and the temporal and analog qualities of biological neurons and synapses could be mimicked RRAM's (Resistive Random Access Memory) memristors. We also present a 3D interconnection roadmap suitable to integrate neural accelerators. Related to neuromorphic hardware design and architectures, we optimize conventional neural network algorithms like Deep Learning (DL) and Spiking Neural Networks (SNNs) by focussing on their most critical parts in terms of power, performance, and area. All this will be leveraged in use case demonstrators for different

applications that need complex machine-learning algorithms in their mobile devices. All these activities are executed in the TEMPO project aiming to broaden the applicability of integrated neuromorphic hardware by means of technological innovation.

Keywords: Neuromorphic computing, edge processing, spiking neural networks, deep learning, hardware, silicon technologies.

2.1 Mobile Devices Call for Efficient Neuromorphic Computing

Increasingly, edge artificial intelligence and machine-learning algorithms enter our day-to-day products and applications such as smart home assistants with natural-language processing, face-recognition-based security systems or autonomous vehicles. In the coming years, the demand for these increasingly complex computational algorithms will only grow further. At this moment, high-end server parks process the data in the cloud.

However, sending data to the cloud costs energy, latency, and is often not preferred for privacy reasons. As such, the ultimate edge artificial intelligence applications require intelligent energy-efficient local processing.

Realizing such intelligent energy-efficient local processing is however extremely challenging. Neuromorphic computing which is modelled according to the sophisticated nerve network of our human brain is often suggested as key technology to realize such processing. The project ECSEL TEMPO (Technology and hardware for neuromorphic computing) [1] aims to progress towards such processing. TEMPO collaboratively develops technology and hardware platforms leveraging emerging memory technologies for neuromorphic computing. The goal is to develop a new way to support a diversity of applications in mobile devices that need complex machine- learning algorithms.

2.2 Neuromorphic Hardware Enables Next Generation AI

Neuromorphic engineering is a ground-breaking approach to the design of computing technology that draws inspiration from the powerful and efficient biological neural processing systems. Neuromorphic devices can carry out sensing, processing, and motor control strategies with ultra-low power performance. Today's neuromorphic community in Europe is leading the State-of-the-Art in this domain. The community counts an increasing number of labs that work on theory, modelling, and implementation of

neuromorphic computing systems using both conventional very large-scale integration (VLSI) technologies, emerging memristive devices, photonics, spin-based, and other nano-technological solutions. To enable the uptake of this technology and to match the needs of real-world applications in future products that solve real-world tasks in industry, healthcare, assistive systems, and consumer devices, extensive work is needed in terms of neuromorphic algorithms, emerging technologies, hardware design and neuromorphic applications respectively.

In the TEMPO project, we consider "neuromorphic" as brain-inspired algorithms, and we focus specifically on conventional DL and SNNs. That way, it is ensured that both established paradigms are covered in the greater domain of brain-inspired computation. Given the slowdown of silicon-only scaling, it is important to extend the roadmap of neuromorphic implementations by leveraging fitting technology innovations. Along these lines, TEMPO sweep technology options, covering emerging memories and 3D integration, and attempt to pair them with contemporary DL and exploratory (SNN) neuromorphic computing paradigms.

Terms like Artificial Intelligence (AI) and Machine Learning (ML) enjoy a popularity trend that is fuelled by a wide variety of applications. They come in a wide variety of underlying algorithms. Regardless of the algorithm, the goal of TEMPO is to implement accurate classifiers and/or predictors of raw data that is either available in a pre-stored location or entering as a stream (images, audio, video, etc). The local deployment of these algorithms, exactly near the generation of raw data, is identified as one of the main progress directions of the overall AI/ML trend [2], which assists the already growing ecosystem that develops and applies neuromorphic algorithms on an increasing number of end-user applications [3]. This observation is echoed additionally by the increasing percentage of custom chips that are designed, which follow the growing AI/ML trend and execute a wide variety of neuromorphic algorithms [4].

To address this, TEMPO aims to **broaden the applicability of integrated neuromorphic hardware** by improving energy efficiency with emerging memory technologies in novel neuromorphic hardware implementations, and to **develop technology platforms** for emerging semiconductor devices and **demonstrate** them for the **energy efficient hardware implementation of neuromorphic workloads**. To achieve this, TEMPO spreads over three action areas as illustrated in Figure 2.1. These action areas cover (1) the definition and the enablement to develop the emerging technologies, (2) the architectural definition and the related neuromorphic hardware design,

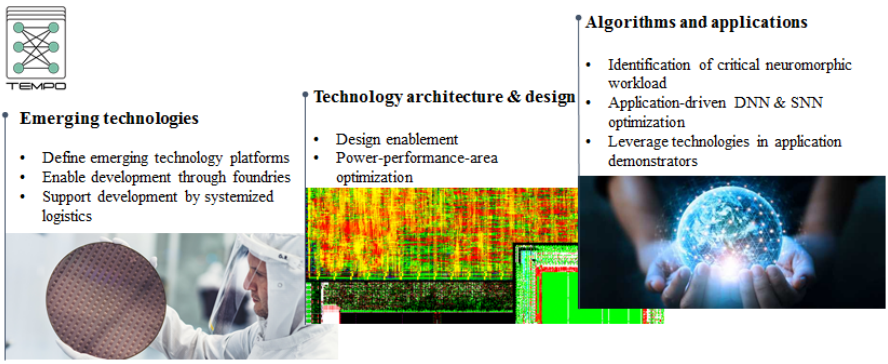

Figure 2.1 TEMPO spreads over three action areas.

and (3) the neuromorphic algorithm design and leverage the neuromorphic technologies for future applications in mobile devices that need complex machine-learning algorithms.

2.3 Building Neuromorphic Hardware

Neuromorphic hardware is the key to sustain the ability of mobile devices to deal with complex machine-learning algorithms. Building such neuromorphic solutions, however, comes with many diverse challenges. These challenges can only be tackled through synergetic collaborations across the entire neuromorphic technology value chain covering major foundries, chip design, system houses, application companies and research partners. TEMPO acts as the umbrella to enable such synergetic activities to address the following objectives:

- **Enable the joint development** of participating European Research and Technology Organisations (RTOs), foundries and leading (application) companies towards the identification of emerging semiconductor technologies that fit best to neuromorphic hardware and address relevant applications indicated by participating end-user partner companies.
- **Evaluate current concepts** for the implementation of neuromorphic hardware according to Key Performance Indicators (KPIs) at the device, architecture and application level, like power consumption, silicon area/cost, latency, throughput, energy for a given application task, memory bottlenecks, manufacturing challenges, operating frameworks.
- **Extend the technology roadmap** that is driven by Integrated Circuits (ICs) designed specifically for AI and ML applications by evaluating

and demonstrating the applicability of emerging technologies that can provide scalable power, performance, and area benefits.

- **Broaden the applicability of neuromorphic hardware**, by designing energy efficient integrated neuromorphic implementations, by fabricating them in collaboration with European foundries and in European cleanrooms, and by benchmarking them in terms of power, performance, and area in the context of pervasive applications that are provided by the end-user partners of the TEMPO project.
- **Exchange wafers** (where applicable) between foundries and the participating RTOs to facilitate the demonstration of functional neuromorphic chips, combining concepts from different RTOs and technologies from industrial companies. This will enable the use of the extensive know-how of European RTOs for future products while maintaining contamination free high-volume manufacturing.
- **Quantify the capability of the most prevalent neuromorphic hardware implementations** by targeting a broad algorithmic spectrum and isolating the critical sections of each algorithm. This includes DL inference (such as CNNs) and SNNs. This wide coverage will result into a Convolutional Neural Network - technology-, design-, and system-aware scorecard containing the most sought-after neuromorphic implementations and their coupling with emerging technologies and applications.
- **Complement existing research** and provide guidance for future directions in the domain of neuromorphic algorithms, design, and systems by assessing the suitability of emerging technologies. The comparative evaluation between implementations of different neuromorphic algorithms can provide guidance to European neuro- morphic research, placing each approach in the context of emerging technologies and relevant applications.
- **Enable the European industry** to remain at the **leading edge** of neuromorphic chip development.

More detailed approaches ang the three action areas defined in TEMPO and illustrated in Figure 1.3.1 are described in the next sections.

2.3.1 Approach to Realise the Emerging Technologies

The core technology component of the TEMPO project is the development of emerging technologies that can provide measurable efficiency benefits to neuromorphic hardware implementations. The objectives with respect to technology are to:

- **Align** the process practices of involved partners, so that base wafers can be optimally exchanged for the development of novel neuromorphic hardware. This includes both the transfer of wafers from the foundries to the involved RTOs and, where/when applicable, the transfer of wafers between the cleanrooms of the RTOs.
- **Match** emerging memory technologies with the proper neuromorphic algorithms, so that hardware integration of the former brings about power, performance, area, and cost benefits.
- **Adjust** process practices so that the integrated emerging memory modules are compatible with traditional semiconductor manufacturing practices.

2.3.2 Approach to Derive the Hardware Architectures and Designs

The core hardware component of the TEMPO project is the development of processing hardware technologies which are efficient to support future AI-intensive mobile applications. The objectives with respect to neuromorphic hardware are to:

- **Develop** novel **architectures** and sub-system designs that help to reduce the memory bottleneck and power consumption, allow for a minimization of required memory space, and minimize the occupied silicon area (i.e., chip cost) while maintaining target accuracy, latency, and throughput.
- **Extend** basic architectures of CNN or SNN arrays with a scalable global communication network to enable high throughput and high complexity applications.
- **Design** modules that use emerging memory technologies to implement the core workloads of the major neuromorphic algorithms.
- **Ensure** component- and system-level compatibility with traditional electronic design flows.
- **Estimate** the power, performance, area, and cost of emerging memory integration for neuromorphic algorithms at the system-on-chip level and compare against contemporary implementations.

2.3.3 Approach Related to Neuromorphic Algorithms and Applications

To put the TEMPO project into the general perspective of accelerated ML, it is fundamental to identify the exact workloads that will be targeted for

efficient and low power hardware integration with advanced technologies. This is a major precondition, as it is of vital importance to optimally concentrate the effort of the project to the fundamental computational bottlenecks identified in the target neuromorphic algorithms. The algorithmic objectives of the TEMPO project are as follows:

- **Profile** target neuromorphic algorithms for computational/memory bottlenecks
- **Identify** the algorithm regions that warrant hardware support
- **Specify** the complexity of the integrated neuromorphic implementations

TEMPO aims to allow applications to make easy use of the new neuromorphic technologies. The objectives to enable this are:

- **Extend the range of applications** to domains requiring (ultra-)high throughput and high complexity such as high throughput imaging, autonomous vehicles, vision enabled robots.
- **Create a demonstration design flow and a tool flow** that connects the target neuromorphic algorithms with the target applications.
- **Prototype the design and tool flows** to illustrate real time characteristics of the target neuromorphic applications, before the emerging technology samples become available.
- **Demonstrate the feasibility** and efficacy of integrated neuromorphic kernels on state-of-the-art benchmarks with functional demonstrators that use or emulate the proposed neuromorphic building blocks.

2.4 Positioning Within the Neuromorphic Computing Landscape

Neuromorphic computing comes in many flavours and forms of maturity. Figure 2.2 gives a simplified but illustrative view of the greater landscape of neuromorphic computing. In terms of implementation, neuromorphic computing can rely in analog, digital or hybrid hardware technologies. In terms of algorithms, the spectrum can range between the compute-intensive deep learning algorithms towards event-based processing like spiking neural network algorithms. The production level maturity is indicatively illustrated in Figure 2.2. Digital processing units like CPU's and GPU's and readily available on the market and are used for compute-intensive tasks in server racks and in the cloud. Commercial solutions are, however, scarcer when considering more analog implementations and/or more transient-based processing. TEMPO covers the complete brain-inspired computation domain,

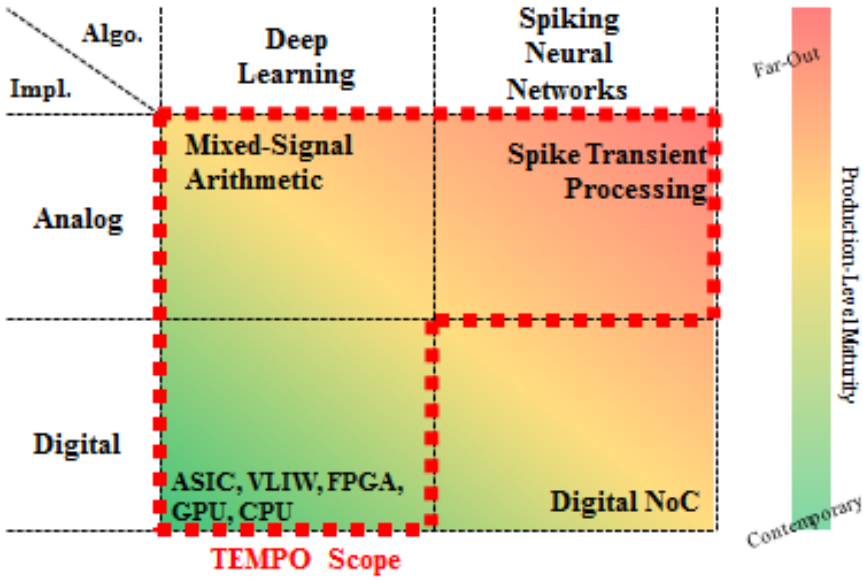

Figure 2.2 TEMPO positioned in the greater landscape of neuromorphic computing.

algorithmically ranging from DL inference engines to exploratory SNNs, and implementation-wise from standard digital to mixed-signal or analog implementations. The quadrant uncovered by TEMPO aims at massively parallel computer architectures. These architectures aim to mimic the implementation of human brains, which are composed of billions of simple computing elements, communicating using unreliable spikes.

The TEMPO project will existing evaluate memory technologies at device, architecture, and application level, and build and expand the technology roadmap for European AI hardware platforms. The project will leverage MRAM, FeRAM and RRAM memory to implement both SNN and Deep Neural Network (DNN) accelerators for 8 different use cases, ranging from consumer electronics to automotive, digital industry and medical applications.

MRAM is a type of memory that stores data magnetically but uses electrons to read and write it. The magnetic character provides non-volatility, which the electronics provides speed. A storage element is comprised of two ferromagnetic layers, consisting of a free layer and a pinned layer, sandwiching a

non-magnetic oxide layer. It works by overcoming the resistance required to switch the magnetization from one direction to the other. Multiple resistance states can be achieved by incorporating domain walls in the free layer. The stochastic nature of switching states in these devices can be employed to mimic the stochastic behavior of synapses.

FeRAM memory uses ferroelectric materials that can switch rapidly between two polarized states. This type of memory offers high performance at low power, along with the added advantage of non- volatility. FeFET (Ferroelectric Field-Effect Transistor) can be tuned to emulate synaptic weight, an important element of neuromorphic computation. One big advantage of FeFET is that some ferroelectric compounds are also Complementary Metal-Oxide Semiconductor (CMOS)-compatible, making it easier to integrate into standard computing platforms. The downside is that the technology also suffers some of the limitations as DRAM (Dynamic Random-Access Memory), including scaling, leakage, and reliability.

RRAM is a form of nonvolatile storage that operates by changing the resistance of a specially formulated solid dielectric material. An RRAM device contains a whose resistance varies when different voltages are imposed across it. RRAM acts as an electronic switch that exhibits non- volatility, i.e., will retain its resistance state even after the voltage is turned off. The main advantages of this memory type are its scalability, CMOS compatibility, low power consumption, and analog conductance modulation. Its suitability for neuromorphic computing is related to the memristor's ability to change its state based on the history of voltages applied to it. As a result of this behaviour, it has the temporal and analog qualities of biological neurons and synapses. However, making these memristors more uniform so they will operate reliably is challenging.

2.5 Targeted Use Cases and Application Domains

The TEMPO project leverages its developed technologies over 8 different use cases over 5 application domains (automotive, food, digital industry, consumer electronics, and medical health). Table 2.1 gives an overview of the different use cases and the related neural network approach and technological choices. The different use cases are driven by the key industry partners within the consortium.

Table 2.1 Edge AI use cases addresses in TEMPO covers five application domains

Use case	Food classification	Traffic object classification	Pattern recognition	Predictive maintenance	Medical image denoising	Lane guidance assistance	Sports assistance	Object recognition
Domain	Food	Automotive	Digital Industry	Digital Industry	Medical health	Automotive	Consumer Electronics	Automotive
Neural Network	DNN/SNN	SNN	SNN	DNN/SNN	SNN	DNN/SNN	DNN/SNN	DNN
FDSOI		Yes		Yes				
Bulk CMOS	Yes		Yes		Yes	Yes	Yes	Yes
Memory type		RRAM		RRAM		FeRAM	MRAM	FeRAM, MRAM
3D SL / stacking	Yes		Yes		Yes	Yes	Yes	Yes

The following sections elaborate some of the envisioned use cases.

2.5.1 Food – Food Classification

This use case focusses on building a network and data pipeline for the classification of western food as illustrated in Figure 2.3. This activity builds a state-of-the-art DNN classifier based on the publicly available dataset Food-101 [5]. The classifier is embedded onto the Edge Tensor Processing Unit (TPU) of Coral [6], which is a low-power DNN accelerator. This will enable to benchmark the developed technology against commercially available hardware solutions.

2.5.2 Automotive – Object Recognition and Sound Localization

This use case focusses on localization and recognition of objects/sound generators. A sound event localization, detection, and tracking network has been developed and could be intended to be on an Field-Programmable Gate Array (FPGA) which emulates the analogue parts of the circuit. A simar demonstrator based on the same principle might be developed by replacing the sound measurements by object visualization through a video camera. Additionally, radar-based object detection might be developed based on hardware developed in the project. Radar has the advantage over video as its network size is considerably smaller.

2.5.3 Digital Industry – Pattern Recognition (Keyword Spotting)

Speech processing enables natural communication with smart phones or smart home assistants. However, continuously performing speech recognition is not energy-efficient and would drain batteries of smart devices. Instead, speech recognition systems passively listen for utterances of certain wake

Figure 2.3 Possible inputs for the western food classification DNN [5].

words to trigger the continuous speech recognition system on demand [8]. In the project, "speech command datasets" have been analysed and features were extracted, and processing pipelines were implemented. The pipelines were used to explore different SNN algorithm approaches. Hybrid variants

will be specified and simulated. After the hybrid variants are evaluated, the algorithms will be integrated into full SNNs.

2.5.4 Consumer – Coaching Biomechanical Assistance (Running)

This use case focusses on real-time running coaching. From an optimized database infrastructure of runners' user data and an improved classification neural networks will be trained. New software that will facilitate broader data and image assimilation from users and classification will be developed of additional input parameters.

2.5.5 Medical Health – Medical Image Denoising

Efficient medical image denoising is essential on mobile X-ray systems. To facilitate this, dataset specification and analysis of the noise characteristics are being made. This shows to be essential and challenging as part of the noise is signal-dependent. Metrics are being proposed to measure and quantify image quality comparisons, and specifications are set for the test cases to be performed on the SNN implementations.

2.6 Neuromorphic Hardware Technologies Being Developed

The developments in TEMPO are still ongoing; it is planned leverage the developed hardware and application results into the envisioned use case applications and related demonstrators by the end of 2022.

The project started with the process technology pathfinding work to enable neuromorphic and AI applications to leverage embedded non- volatile memories (eNVMs). This pathfinding work included the design of process technology test vehicles and process flows. At the same time, core building blocks and accelerator architectures have been designed to leverage the memory technologies in the application demonstrators. Basic neuromorphic building blocks were investigated with a focus on the development of neuromorphic–ready NVM blocks, the modelling and simulation of eNVM, the quantification of the technology features and neuromorphic implementation of eNVM. 3D specifications suited for DNN accelerators have been defined and a design flow to be able to quantify performance and energy impact of 3D interconnect has been set-up. Design and architecture exploration, specification, and design of critical building blocks to enable full accelerator IP blocks has been done.

Later in the project, the first hardware and algorithms were leveraged towards the applications via the different use cases. In the domain of **emerging technologies**, basic neuromorphic building blocks (MRAM, Oxide Random Access Memory - OxRAM and FeFET) were investigated, with a focus the development of neuromorphic–ready NVM blocks, modelling and simulation of eNVM, and the quantification of the technology features and neuromorphic implementation of eNVM. Also features of embedded memory for Neuromorphic Accelerators have been investigated, such as multi-level memory and the synapticity/plasticity of the memories. In the domain of **technology integration**, compact models were created based on the data from first OxRAM, Phase-Change Random Access Memory (PCRAM) and FeFET implementations. Also, 3D specifications suited for DNN accelerators have been defined and the 3D place and route (PnR) design flow has been created to quantify performance and energy impact of the 3D interconnects. An illustration of an envisioned 3D interconnect roadmap suitable for typical neural accelerators is illustrated in Figure 2.4. In the domain of **neuromorphic hardware design and architectures**, potential design, and architectures of the most critical neuromorphic DNN and SNN building blocks in terms of power, performance and area have been explored. Finally, in the domain of **application specification and demonstration**, the use cases and related data sets have been defined and the reference platform has been chosen and benchmarked. These uses cases have been elaborated in section 2.5. Theses use cases are being implemented towards demonstration.

TEMPO will continue to combine both the developed hardware and application results to enable demonstration of energy efficient accelerators for the different use cases defined in the project.

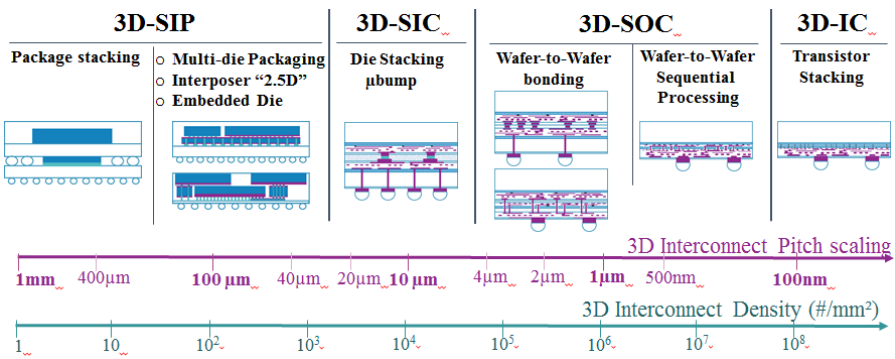

Figure 2.4 3D landscape, ordering of 3D technologies according to the system-level wiring hierarchy [11, 12].

2.7 Conclusion

In most application domains, the amount of data produced in sensors and devices is exploding. Sending this data to the cloud costs energy, latency, and is often not preferred for privacy reasons. Applications relying on artificial intelligence in the edge require intelligent energy-efficient local processing. The TEMPO project develops such energy efficiency neuromorphic hardware with emerging memory technologies like MRAM, FeRAM and RRAM, and develops technology platforms for emerging semiconductor devices. In the domain of emerging technologies, the project investigated the different memory types to confirm their suitability and limitations towards offering the needed neuromorphic features and implementation. Compact models were created based on the first memory implementations and a 3D interconnect roadmap suitable for typical neural accelerators has been designed and presented. To enable neuromorphic hardware design, the architecture of the most critical neuromorphic DNN and SNN building blocks have been explored in terms of power, performance, and area. This paves the way to demonstrate these technologies for the neuromorphic workloads required in the envisioned use cases. These use cases and their dataset requirements have been specified as discussed in this article. These use cases cover a broad range of application fields within automotive, consumer electronics, digital industry, food, and medical health. As such, the TEMPO project is successfully pursuing its goal to broaden the applicability of integrated neuromorphic hardware.

Acknowledgements

TEMPO (Technology & hardware for nEuromorphic coMPuting) is a European innovation project. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme, and from Belgium, France, Germany, The Netherlands, and Switzerland. TEMPO was kicked off on the 1st of April 2019 and has a duration of three years. The consortium of this ambitious project consists of eighteen members from five different European countries. Imec takes the lead as the sole Belgian consortium partner. The other consortium members are, for France: CEA-Leti, ST-Microelectronics Crolles, ST-Microelectronics Grenoble and Thales Alenia Space. For Germany: Bosch, Fraunhofer (EMFT, IIS, IPMS), Infineon, Valeo, Innosent, TU Dresden and Videantis. For the Netherlands: imec the Netherlands, Philips Electronics, Philips Medical Systems and Ato-gear. For Switzerland: aiCTX and the

University of Zürich. For more information on the project and the consortium partners: https://tempo-ecsel.eu/

References

[1] TEMPO project. Technology and Hardware for Neuromorphic Computing. Available online at: https://tempo-ecsel.eu/
[2] Deloitte, "Technology, Media and Telecommunications Predictions (TMT) 2021. Available online at: www2.deloitte.com/be/en/pages/technology-media-and-telecommunications/articles/tmt-predictions.html.
[3] NVIDIA, "Annual investor day", 12 April 2021. Available online at: https://investor.nvidia.com/events-and-presentations/events-and-presentations/event-details/2021/NVIDIA-Annual-Investor-Day/default.aspx.
[4] Deloitte, "Hitting the accelerator: the next generation of machine- learning chips", 2017. Available online at: www2.deloitte.com/content/dam/Deloitte/global/Images/infographics/technologymediatelecommunications/gx-deloitte-tmt-2018-nextgen-machine-learning-report.pdf.
[5] TensorFlow, "TensorFlow datasets: a collection of ready-to-use datasets – dataset Food-101", 2021. Available online at: www.tensorflow.org/datasets/catalog/food101.
[6] Coral, "Products", 2021. Available online at: https://coral.ai/products/.
[7] Robert Bosch, "Embedded siren detection", 2021. Available online at: www.bosch.com/stories/embedded-siren-detection.
[8] S. Mittermaier, L. Kürzinger, B. Waschneck, G. Rigoll, "Small- Footprint Keyword Spotting on Raw Audio Data with Sinc- Convolutions", arXiv, 1911.02086, 2020. Available online at: https://arxiv.org/abs/1911.02086.
[9] Ato-Gear, "Arion", 2021. Available online at: www.arion.run.
[10] Philips, "Mobile digital radiography system", 2021. Available online at: www.philips.nl/healthcare/product/HC712001/mobilediagnost-wdr-mobile-digital-radiography-system.
[11] imec, "A 3D technology toolbox in support of system-technology co-optimization", 2019. Available online at: www.imec-int.com/en/imec-magazine/imec-magazine-july-2019/a-3d-technology-toolbox-in-support-of-system-technology-co-optimization.
[12] Samavedam, S. M., et al., "Future Logic Scaling: Towards Atomic Channels and Deconstructed Chips," IEEE International Electron Devices Meeting (IEDM), 2020. https://doi.org/10.1109/IEDM13553.2020.9372023