NimbleAI: Towards Neuromorphic Sensing-Processing 3D-integrated Chips

Xabier Iturbe, Nassim Abderrahmane, Jaume Abella, Sergi Alcaide, Eric Beyne, Henri-Pierre Charles, Christelle Charpin-Nicolle, Lars Chittka, Angélica Dávila, Arne Erdmann, Carles Estrada, Ander Fernández, Anna Fontanelli, José Flich, Gianluca Furano, Alejandro Hernán Gloriani, Erik Isusquiza, Radu Grosu, Carles Hernández, Daniele Ielmini, David Jackson, Maha Kooli, Nicola Lepri, Bernabé Linares-Barranco, Jean-Loup Lachese, Eric Laurent, Menno Lindwer, Frank Linsenmaier, Mikel Luján, Karel Masařík, Nele Mentens, Orlando Moreira, Chinmay Nawghane, Luca Peres, Jean-Philippe Noel, Arash Pourtaherian, Christoph Posch, Peter Priller, Zdenek Prikryl, Felix Resch, Oliver Rhodes, Todor Stefanov, Moritz Storring, Michele Taliercio, Rafael Tornero, Marcel van de Burgwal, Geert van der Plas, Elisa Vianello, and Pavel Zaykov

***Abstract*—The *NimbleAI* Horizon Europe project leverages key principles of energy-efficient visual sensing and processing in biological eyes and brains, and harnesses the latest advances in 3D stacked silicon integration, to create an integral sensing-processing neuromorphic architecture that efficiently and accurately runs computer vision algorithms in area-constrained endpoint chips. The rationale behind the NimbleAI architecture is: sense data only with high information value and discard data as soon as they are found not to be useful for the application (in a given context). The NimbleAI sensing-processing architecture is to be specialized after-deployment by tunning system-level trade-offs for each particular computer vision algorithm and deployment environment. The objectives of NimbleAI are: (1) 100x performance per mW gains compared to state-of-the-practice solutions (i.e., CPU/GPUs processing frame-based video); (2) 50x processing latency reduction compared to CPU/GPUs; (3) energy consumption in the order of tens of mWs; and (4) silicon area of approx. 50 mm^2.**

***Index Terms*—Neuromorphic, computer vision, 3D silicon, event-based vision, in-memory computing, eFPGA, RISC-V, virtual neural networks, light-field vision, online learning**

X. Iturbe, C. Estrada, A. Fernández and A. Dávila are with IKERLAN, Basque Country (Spain); C. Hernandez, R. Tornero and J. Flich are with Universitat Politecnica de Valencia, Spain; J. Abella and S. Alcaide are with Barcelona Supercomputing Center (BSC), Catalonia (Spain); F. Resch and R. Grosu are with TU Wien, Austria; G. Van der Plas, M. van de Burgwal, M. Storring, C. Nawghane and E. Beyne are with IMEC, Belgium; A. Erdmann is with Raytrix, Germany; N. Abderrahmane, J.L. Lachese and E. Laurent are with MENTA, France; N. Mentens and T. Stefanov are with Universiteit Leiden, Netherlands; P. Zaykov, Z. Prikryl and K. Masařík are with CODASIP, Czech Republic; M. Lindwer, O. Moreira and A. Pourtaherian are with GrAI Matter Labs (GML), Netherlands; N. Lepri and D. Ielmini are with Politecnico Milano, Italy; O. Rhodes, M. Luján, L. Peres and D. Jackson are with University of Manchester, UK; B. Linares-Barranco is with CSIC, Spain; A. Fontanelli and M. Taliercio are with Monozukuri (MZ Technologies), Italy; M. Kooli, H.P. Charles, J.P. Noel are with CEA-LIST, University Grenoble Alpes, France; C. Charpin-Nicolle and E. Vianello are with CEA-LETI, Univ. Grenoble Alpes, France; P. Priller is with AVL List, Austria; E. Isusquiza is with ULMA Medical Technologies, Basque Country (Spain); L. Chittka is with Queen Mary University of London, UK; A.H. Gloriani and F. Linsenmaier are with Viewpointsystem, Austria; C. Posch is with PROPHESEE, France; G. Furano is with ESA ESTEC, Netherlands. All co-authors are listed in alphabetical order except for the main and corresponding author (xiturbe@ikerlan.es).

I. Introduction

CPU/GPU-based computer vision solutions are very inefficient compared to biological eye-brain visual systems, which are honed by natural selection and apply the fundamental energy-saving principle of capturing, processing and storing data only when necessary. Hence, eyes continuously sense and encode the changing surrounding environment in a way that is manageable for the brain.

The recently started NimbleAI project leverages key principles of energy-efficient light detection in eyes and visual information processing in brains to create an integral sensing-processing neuromorphic chip that adopts the biological data economy principle at different system levels, and builds upon the latest advances in 3D stacked silicon integration. NimbleAI aims to deliver two world's firsts: (1) a light-field dynamic vision sensor for monocular image-based depth perception; and (2) an event-driven end-to-end perception stack ('*visual pathway*') that runs industry standard Convolutional Neural Networks (CNNs). Since manufacturing a full 3D testchip is prohibitively expensive, NimbleAI will prototype key components via small-scale 2D stand-alone testchips. This cost-effective use of silicon is expected to allow us to produce high confidence research conclusions and silicon-proven neuromorphic IP.

This article discusses the main functioning principles of the NimbleAI architecture and existing system-level trade-offs: (1) sense only significant light changes (visual events) at the optimal spatio-temporal resolution; (2) distill sensed visual events to increase information-efficiency; (3) process selected information-rich events using minimal energy at the optimal DVFS point; and (4) route event-flows across the 3D stacked sensing-processing architecture to minimize data movement along shortest physical paths. The article is organized as follows. Section II introduces the major challenges of AI-enabling technologies that are addressed by NimbleAI. Section III outlines the overall NimbleAI concept and section IV describes the proposed architecture. Finally, Section V sums up the main takeaways to conclude the paper.

 http://dx.doi.org/10.23919/DATE56975.2023.10136952

II. Current context and challenges

NimbleAI deals with four main challenges and limitations of current computer vision algorithms and AI hardware.

C1.- Complexity of AI models: Accuracy of computer vision algorithms is commonly opposed to efficiency. CNNs are typically scaled up to increase accuracy by adding more layers or by enlarging these to process images at a higher resolution. On the other hand, state-of-the-practice edge CNNs typically rely on downscaling the resolution of full input images to keep workloads manageable by current inefficient processing architectures (see C3), thus sacrificing accuracy. Inaccuracies become greater when shrinking large industry standard CNNs to fit in resource-constrained edge and endpoint devices.

C2.- Performance and latency: State-of-the-practice computer vision systems are frame-based, which means that they periodically acquire and process full-size images in a layer-after-layer mode. Hence, the computation of one layer must be completed on the whole frame before the computation of the next layer starts. This results in growing inference delays as algorithms include more layers and sensor resolution increases.

C3.- Energy-efficiency of processor architectures: The current state-of-the-practice processor landscape includes general-purpose (CPU/GPU) and AI-specialized (NPU/TPU) architectures. CPU/GPUs are largely inefficient due to the continuous back-and-forth transfers of data (and instructions) with memory, whereas efficiency improvements brought about by NPUs (Neural Processing Units) and TPUs (Tensor Processing Units) depend to a great extent on the ability of the host CPU to split AI processing into matrix operations of similar dimensions to those for which the NPU/TPU architecture was optimized. State-of-the-art neuromorphic architectures, on the other hand, implement brain-inspired (event-driven) neural networks to enormously increase energy-efficiency as they process only changes in their inputs [1]. Yet, only a few neuromorphic architectures promise to meet the high energy-efficiency levels with low energy budgets required at the endpoint (e.g., Innatera, SynSense, GrAI Matter Labs - GML, etc.). An important limitation of neuromorphic chips is that the size of neural networks that can run is restricted by the implemented neuron count in silicon. Innatera and Synsense commercial chips implement only 1,000 neurons, greatly limiting their use to one dimensional applications such as audio. On the other hand, TrueNorth is the largest chip that IBM has ever built: at 500 mm^2 can hold only 1 M neurons [2], while real-world (image) applications typically require 10-20 M neurons and endpoint chips are typically 50 mm^2.

C4.- System integration: CPU/GPUs and NPU/TPUs are not typically integrated such that they can seamlessly and efficiently process data streams from sensors or interface to pre- and post-processing kernels. For example, TPUs do not have image sensor interfaces and hence need to rely on a host processor to capture and transmit video sequences to the TPU engine. For each video frame, this process may take factors more time than the TPU's actual AI processing of that same frame. Similar constraints hold for GPU and NPUs.

III. The NimbleAI Concept

NimbleAI considers that processing begins in the sensor. In fact, important efficiency gains are expected from the use of in-sensor analog logic and novel dynamic vision sensing concepts that will be investigated for the first time in this project. These concepts include: (1) *digital-foveation* to dynamically allocate sensing resolution based on the information value brought about by each sensor region; and (2) coupling of light-field microlenses with the Dynamic Vision Sensor (DVS) to enable *event-based light-field perception*.

NimbleAI will study techniques to capture and optimally represent the spatio-temporal evolution of 3D scenes using minimal visual event-flows that match the optimization features implemented in the downstream processing and inference engines, and thus reduce energy consumption and latency of the whole architecture. The expectation is that by investing some computing power and energy to gain some situational awareness early, a major reduction of the amount of data to be processed will be achieved, saving lots of energy by doing that. Early perception in NimbleAI is inspired by unconscious visual processing and neural signalling in biological systems, and hence will be largely invisible to the user application yet adjustable through user-driven directives. With this, we expect to increase the amount of meaningful information that can be obtained as DVS resolution scales up, which is a major open challenge in event-based vision [3].

As shown in Fig. 1, one of the novel system-level bio-inspired concepts that will be explored in the project are event-driven *visual pathways* for optimal sensing and processing of feature-rich regions of interest (ROIs). As opposed to current event-based vision approaches that are yet limited (e.g., [4]), NimbleAI aims to demonstrate event-driven end-to-end visual pathways that can run industry standard AI models such as CNNs. Visual pathways will be assigned to ROIs in a one-to-one fashion: each pathway will span the assigned ROI sensor area and will use dedicated Through Silicon Viases (TSVs) to downstream visual events to the processing and inference engines in the interior layers of the 3D stacked architecture.

We pose that visual pathways are an elegant way to answer challenge C4 and harness the increased bandwidth brought about by 3D integration, taking advantage of the irregular distribution of visual information and uneven temporal dynamics in the scene. In fact, each visual pathway will be configured (and optimized) independently and dynamically, from sensor to processing, at the accuracy (e.g., digital foveation) and latency levels determined for that ROI based on its dynamics and information value. Conversion of visual events delivered by the DVS to neural events in the inference engine will be switched between time-driven and performance-driven, and the inference engine will be accordingly adjusted to work at the optimal point to serve the workload expected at each time.

NimbleAI envisions a two-stage inference approach, where the two stages will reinforce each other to perform more efficiently as the deployment environments become more familiar and visual stimuli are better understood.

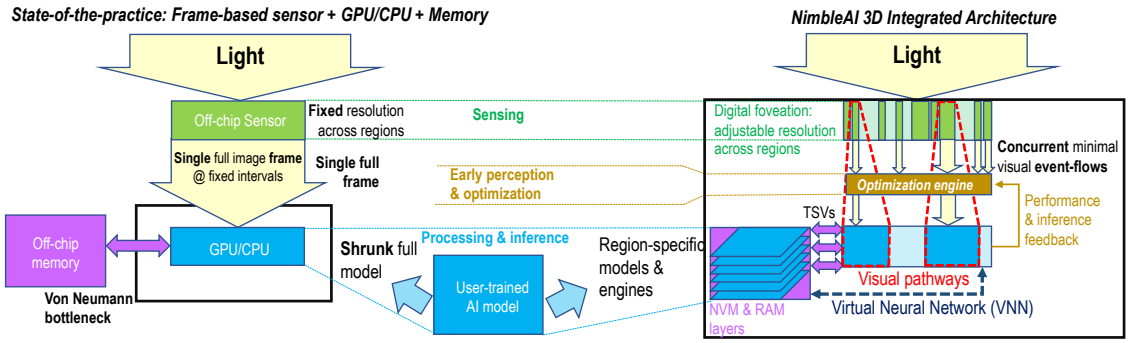

Fig. 1: State-of-the-practice CPU/GPU (left) vs NimbleAI approach (right).

1.- An always-on *early perception and optimization stage* (see section IV.B) implements selective attention algorithms to identify ROIs and configure accordingly the visual pathways. This includes selecting the most appropriate sensor resolution for each ROI and routing sensed visual events to the most appropriate processing kernel and AI model (e.g., CNN) running on downstream engines for efficient end-to-end region inference. NimbleAI will explore ultra-low energy and low-latency advantages of Spiking Neural Networks (SNNs) to power the early perception and optimization stage, as well as energy-efficient online learning rules to achieve specialization in dealing with deployment-specific visual stimuli. Hence, optimization SNNs will receive inference feedback from user models to continuously improve on dynamically selecting ROIs. This can be seen as a partial knowledge transfer from user-trained models to the early perception stage to optimize overall functioning. Hence, energy consumed to complete end-to-end inference also serves the purpose of adjusting the energy-saving mechanism in the early perception stage. SNNs will also receive performance feedback and voltage/temperature (V&T) analytics from monitors embedded along the visual pathways to continuously learn how to tune the execution conditions to be more efficient. This includes finding and adjusting the inference engine optimal working point at each time and context.

2.- An *inference stage* (see section IV.C) implements pre- and post-processing kernels on the downstream processing engine and runs industry standard CNNs on the event-driven dataflow inference engine. Processing in these components will be on-demand and optimized for the specific characteristics of each ROI. To deal with challenge C3, NimbleAI will explore the novel concept of *Virtual Neural Networks (VNNs)* to allow users to run large and accurate event-driven inference models in only 50 mm^2 chips. As shown in Fig. 1, this concept will be supported by dedicated TSVs and 3D layers of RAM and NVM that will be architected to create a high-bandwidth and high-density memory hierarchy for quickly swapping active and non-active neurons and (parts of) CNNs in the event-driven dataflow inference engine.

Neuromorphic event- and region-based sensing and processing in NimbleAI will help limit the complexity and energy-consumption of AI models, and thus deal with challenges C1 and C2. AI models and algorithms that work on selected image regions are simpler than those that work on full images, and event-driven networks that execute on neuromorphic hardware only consume energy when there are significant changes in their neuron states, which are themselves triggered only by significant changes in sensed visual data. Hence, as opposed to state-of-the-practice, which downscale the resolution of input images to keep workloads manageable, NimbleAI will process selected full-resolution image parts for better accuracy. Also, as opposed to state-of-the-practice approaches, where more complex/accurate AI models translate directly into more computing and energy consumption, in NimbleAI model complexity to workload translation will be dynamically regulated through runtime optimization mechanisms that control visual event generation and processing rates along visual pathways.

This unique optimization approach is opposed to the current situation in which performance and accuracy trade-offs are often presented to users as a necessity at the design phase that remains fixed in deployment. NimbleAI will not oblige users to choose between accuracy or efficiency. Instead, it will offer to the user a number of system-level runtime optimization strategies that will be continuously refined by means of online learning and applied directly on the user-trained models.

IV. The NimbleAI 3D stacked architecture

This section describes each of the stacked layers in the 3D NimbleAI chip shown in Fig. 2.

A. Light-field DVS with digital foveation

NimbleAI will implement a digital foveation mechanism to dynamically group and ungroup DVS pixels in the sensor layer to form macro-pixels with varying resolution levels across image regions based on the information value each region brings to the application. If the selective attention algorithms (subsection IV.B) identify something potentially meaningful, DVS macro-pixels in that region will be ungrouped to form a foveated full-resolution ROI that will be processed by a dedicated downstream inference engine. Several foveated regions that match the size, shape, resolution and moving dynamics of the recognized and tracked objects in the scene could be sensed simultaneously to achieve the most accurate results without unnecessarily increasing the amount of data to be processed.

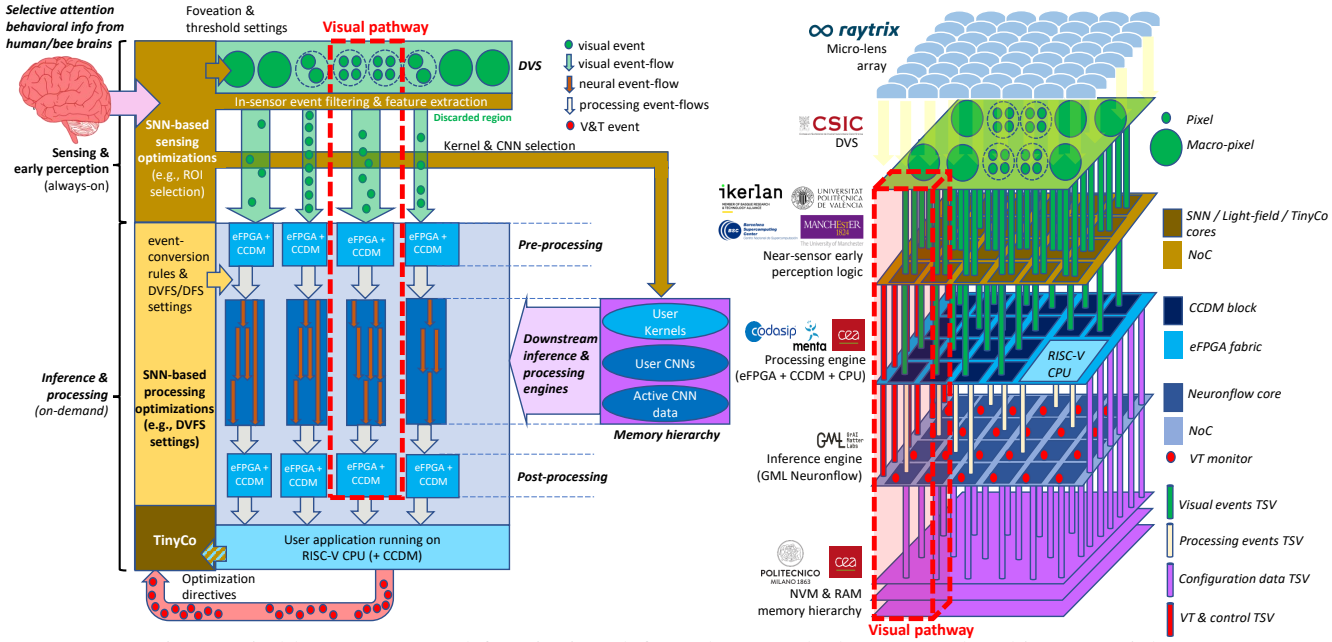

Fig. 2: NimbleAI: conceptual functioning (left) and 3D stacked conceptual architecture (right).

NimbleAI will also be looking at insect compound eyes to enable 3D perception for accurate depth and motion estimation. Namely, the project pursues to adapt Raytrix light-field technology (i.e., micro-lenses) [5] and algorithms to encode 3D visual scenes in the form of sparse events that also include depth information: (x,y,z,t). The fact that DVS events reflect moving edges of the objects in the scene and that light-field algorithms rely on correlations between neighbour data with lots of redundancy, leads us to think that large amounts of processing and energy could be saved by combining both technologies. In fact, it has already been demonstrated that boundaries-first processing lends well to DVS events while putting less pressure on the hardware [7]. Depth information will open new opportunities for improving both event selection in the early perception stage (e.g., ROIs: nearby objects) and perception accuracy in the inference stage.

NimbleAI will investigate 3D silicon integration to vertically stack the additional in-sensor logic needed to implement the functionalities described above, and CEA RRAM [6] to store DVS adjustable parameters (e.g., calibration words and thresholds), with the objective of reducing the impact on the pixel footprint and thus support sensor resolution scaling. NimbleAI will manufacture a DVS testchip with limited resolution and using affordable technology nodes to demonstrate the digital foveation and adaptive region-based sensing mechanisms with adjustable parameters. To demonstrate 3D perception, a light-field-enabled DVS prototype will be manufactured coupling a custom-made array of micro-lenses designed by Raytrix on a commercial PROPHESEE sensor.

B. Near-sensor early perception and optimization

NimbleAI will rely on a tiny controller (TinyCo) to manage sensing and processing in visual pathways. The TinyCo can be configured to provide full control of sensing and processing to the user application (top-down decision-making), or to make optimization decisions autonomously (bottom-up decision-making) for improved performance. In the former case, for example, the user can explicitly configure ROI limits, whereas in the latter case the user can provide simple rules and thresholds to support autonomous decision-making. For example, event density thresholds in the spatio-temporal domain to remove unwanted visual events such as noise.

To support autonomous bottom-up decision-making in the TinyCo, NimbleAI will explore uses of SNN algorithms witch rich temporal dynamics to run ultra-energy-efficient near-sensor visual scene analysis and visual attention. The objective of SNN is to provide initial feature extraction and delimit ROIs of almost arbitrary size around collections of identified key features. SNN-based optical flow processing will also be investigated to estimate the speed and direction of movement in the scene. Movement direction estimations will help drive digital foveation in the DVS, whereas speed estimations will help decide the optimal time interval to accumulate visual events prior to be sent to the inference engine. SNNs are expected to respond rapidly to dynamic changes in the visual input as visual events will trigger SNN-based processing.

It has already been demonstrated that complex visual-cognitive tasks performed by bees can be modelled with SNN models [8]. Following these findings, NimbleAI will explore models and topologies of comprehensive SNN-like circuits of brains in bees to assess how internally generated oscillations combined with sensory visual events could generate useful types of attentional performance. Likewise, modular and scalable SNN topologies will be investigated to extract optical flow from light-field visual event-flows, analyzing the relations between the size of micro-lenses, DVS pixels (and macro-pixels) and the number of SNN neurons (see section IV.A).

One major objective of NimbleAI is to achieve ultra-energy efficiency by specializing processing to deployment-specific

visual events. This involves optimizing (or refining) processing of (new) visual inputs using pre-trained neural networks. In this regard, SNNs are particularly well suited to online training, as their event-based learning rules typically use only information local to the synapse, requiring significantly less computing than the error back-propagation techniques employed to train traditional artificial neural networks [9]. While the focus will be on energy-efficient inference, NimbleAI will also experiment with a range of synaptic plasticity mechanisms such as reinforcement learning and neuromodulation techniques [10], to explore how meaningful visual behaviour can be adjusted on-the-fly, using reward/punishment signals from other parts of the system; e.g., feedback correct/incorrect selection of ROIs from downstream inference engines.

SNNs will also be explored to make inference and processing related optimization decisions. In fact, the rich temporal dynamics of SNNs are expected to allow them to harness visual continuity in the scene and make more accurate workload evolution predictions. At design time, SNNs will be trained using performance and integrated circuit implementation information, including: (1) energy-performance curves of inference CNNs with different DVFS settings, (2) energy-latency-accuracy curves for different sensor and inference engine settings, and (3) location-dependent thermal dissipation and transmission characteristics in the 3D architecture. At runtime, SNNs will use real-time analytics delivered by activity and V&T monitors as feedback for online learning.

Following the same philosophy as for DVS, NimbleAI will design on-chip V&T monitors that generate digital events when they detect voltage and/or temperature variations above or below configurable thresholds. These events will be directly fed into the SNN to take advantage of event-based low-latency processing. SNN-based processing might be especially relevant in next-generation ultra fine-grain DVFS systems to approach brain-like self-regulated energy distribution mechanisms; i.e., anticipate energy needs across regions. Although this concept might have a longer-term impact, we think that providing SNNs with a unified event-based view of both external (i.e., visual) and internal (i.e., activity and V&T) insights is a very interesting approach to explore holistic optimization decisions that encompass both sensing and processing.

NimbleAI will rely on using SNN software such as NEST/NEURON to carry out model and topology exploration, as well as neuromorphic hardware such as SpiNNaker [10] and commercial spiking-based chips to test the selected models and topologies in real-time applications.

C. Inference and processing

NimbleAI will leverage state-of-the-art event-driven dataflow architectures (i.e., GML NeuronFlow [11]) as main inference downstream engine. As it occurs with SNNs, the type of (neural) events that are processed by event-driven dataflow architectures correspond accurately with DVS (visual) events, thus maximizing end-to-end efficiency along visual pathways. Recent research has shown that industry standard CNNs designed and trained with popular AI frameworks (e.g., TensorFlow) can be converted to equally accurate event-driven networks with lower computational complexity and hence greater energy-efficiency [12].

Spatial, temporal and neural activation sparsity will be effectively exploited in the event-driven inference engine to improve energy-efficiency and reduce latency. Hence, compute and event propagation only occurs on sufficiently significant neuron state changes, which have not been filtered out because of sparsity. The processing through the dataflow engine proceeds in a systolic array manner, forming "waves" that flow outwards from physical entry points from 3D stacked layers. To support visual pathways and optimally benefit from the DVS foveation approach, the inference engine will be able to run multiple CNNs simultaneously to which visual event-flows are streamed to. At each time, only CNNs that match data resolution and temporal dynamics of visual stimuli detected in ROIs will be active.

The NimbleAI inference engine will be enabled for running large VNNs with improved accuracy on resource- and area-constrained 50 mm^2 chips. VNNs will be supported by efficient hardware mechanisms to virtually augment the effective count of neurons integrated in limited chip silicon area. This will be achieved by enabling store and restore network parameters and data on a 3D memory hierarchy that includes stacked high-density RRAM and low-access time RAM layers. The latter memory hierarchy will implement prefetching and synchronization mechanisms integrated within the neuron processing pipelines to ensure that neural network parameters and data are accessed and deployed in a timely and efficient manner. Moreover, novel techniques to support compressed synaptic weights, connectivity, and state encoding and storage will be explored to reduce overall data movement. To achieve ultra-high RRAM capacity, NimbleAI will explore and design 3D crosspoint arrays with one-selector/one-resistor (1S1R) memory architectures [13].

Accompanying the inference engine and VNN-supporting memory layers, the NimbleAI architecture will include one (or several) processing engines consisting of a CODASIP RISC-V extensible CPU and a MENTA eFPGA fabric for hosting DSP-like pre- and post-processing kernels. Besides application-specific processing, this engine will adapt the format and properties of incoming visual event-flows to exploit the hardware optimization mechanisms implemented in the event-driven dataflow inference engine (e.g., sparsity exposure).

RISC-V CPU and eFPGA fabric will integrate CEA in-memory computing Computational SRAM (CSRAM) [14] blocks to exploit vector computation with less data movement. Furthermore, coupling the CSRAM with eFPGA results in Closely Coupled DSP-Memory (CCDM) blocks that provide data parallelism at various granularities to deal with data-intensive operation patterns. Likewise, coupling eFPGA and CCDM with CPU will allow an existing processor design to be specialized for application-specific processing by adding custom instructions (e.g., vector processing and ad-hoc multiply and accumulate) and microarchitecture features, even after deployment. This integrated adaptable processing architecture

will reduce data traffic between the CPU and the memory by performing logic, arithmetic, and DSP operations directly in-memory using CCDM. NimbleAI will study the programming model for such an integrated processing architecture that includes CCDM, eFPGA and CPU, and will specify the instruction format generated by the CPU that ultimately defines the user application code. This programming model will be implemented and integrated within the HybroGen software environment for compilation and code generation [15].

D. Physical structure and implementation

A major objective of NimbleAI is to integrate the components explained in previous subsections into an optimized 3D stacked silicon architecture, where each layer is to be implemented using the most appropriate process technology.

To achieve this, the project will develop an EDA tool that supports novel co-design methodologies covering technology-aware 3D architecture exploration across layers and integration to physical implementation. The NimbleAI 3D EDA tool will build upon MZ Technologies Genio 3D tool and third-party physical implementation and signoff EDA tools for 2D IC design. The architecture exploration will consider technology-related aspects, such as process technology and component size trade-offs, and will help make decisions related to layer floor-planning, vertical arrangement of layers, and inter-layer TSV locations to increase computation density and performance as well as boost communication bandwidth and energy-efficiency. A special focus will be put in designing thermal models of the 3D architecture to pinpoint the locations where to insert V&T monitors to increase the visibility of energy dynamics and thermal dissipation that guide the runtime optimization decisions. Thermal models will be developed using commercial finite element software Marc. The 3D architecture along with TSVs will be converted to a compact model to be consulted by the NimbleAI 3D EDA tool.

NimbleAI will be looking into integrating IMEC's latest generation of baseline TSV models in a suitable format for the pathfinding engine in the 3D EDA tool to ensure compatibility with commercial silicon technologies, including hardware validation of TSV processes. The greatest density of TSVs is expected to connect the DVS sensor with the near-sensor logic layer and the processing engine. The lowest density of TSVs is expected to connect the near-sensor logic layer to the processing and inference engines, as well as to support control and monitoring data exchanges among layers. Finally, a medium density of TSVs is expected to support the VNN mechanism, connecting the Neuronflow cores in the inference engine with the memory layers. Thermal feasibility and TSV integration limitations will be studied in all these cases.

V. Takeaways

NimbleAI takes inspiration from ultra-energy-efficient eye-brain systems, even combining divergent evolutionary developments, such as foveation in vertebrate eyes and compound insect eyes. The project expects to deliver 100x energy-efficiency improvement and 50x latency reduction (w.r.t. CPU/GPUs processing frame-based video) by using: (1) DVS sensing with digital foveation and selective attention; (2) event-driven visual inference at optimal DVFS point; (3) specialized processing with in-memory computing; (4) 3D-integrated visual pathways; and (5) system-level optimizations to continuously adjust sensing and processing in each visual pathway to operate jointly at the optimal temporal and data resolution.

NimbleAI will design EDA tools to customize and integrate the technologies and mechanisms above on a sensing-processing 3D silicon stacked chip. The project will deliver a prototypic implementation of this 3D architecture using an FPGA, small-scale 2D stand-alone testchips and commercial neuromorphic chips. This prototype will be accompanied by the corresponding programming tools to develop and run computer vision applications on it. It will be flexible to accommodate user application IP and will be aimed for use as a research vehicle to test novel AI algorithms and runtime optimizations in use-cases related to medical imaging, autonomous driving, eye tracking and space missions. It is expected that findings coming out from this research will lead to practical implementations in next-generation commercial chips.

Acknowledgments

NimbleAI has received funding from the EU's Horizon Europe Research and Innovation programme (Grant Agreement 101070679), and by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Agreement 10039070). See: https://www.nimbleai.eu.

References

[1] C.D. Schuman et al., "Opportunities for Neuromorphic Computing Algorithms and Applications," Nature Computational Science, vol. 2, 2022.
[2] P. Merolla et al., "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," SCIENCE, vol. 345, no. 6197, 2014.
[3] D. Gehrig and D. Scaramuzza, "Are High-Resolution Event Cameras Really Needed?," ArXiv abs/2203.14672, 2022.
[4] J. Hagenaars et al., "Self-Supervised Learning of Event-Based Optical Flow with Spiking Neural Networks," Intl. Conf. on Neural Information Processing Systems, 2021.
[5] Ren Ng et al., "Light Field Photography with a Hand-held Plenoptic Camera," Stanford university, 2005.
[6] E. Esmanhotto et al., "High-Density 3D Monolithically Integrated Multiple 1T1R Multi-Level-Cell for Neural Networks," IEEE Intl. Electron Devices Meeting (IEDM), 2020.
[7] C. Kim et al., "Scene Reconstruction from high Spatio-Angular Resolution Light Fields," ACM Transactions on Graphics, vol. 3, no. 4, 2013.
[8] F. Peng and L. Chittka, "A Simple Computational Model of the Bee Mushroom Body Can Explain Seemingly Complex Forms of Olfactory Learning and Memory," Current Biology, vol. 2, no. 2, 2017.
[9] J.L. Lobo et al., "Spiking Neural Networks and Online Learning: An overview and perspectives," Neural Networks, vol. 121, 2020.
[10] Mikaitis et al., "Neuromodulated Synaptic Plasticity on the SpiNNaker Neuromorphic System," Frontiers in Neuroscience, vol. 12, 2018.
[11] O. Moreira et al., "NeuronFlow: a Neuromorphic Processor Architecture for Live AI Applications," Conf. on Design, Automation and Test in Europe, 2020.
[12] L. Deng et al., "Understanding and Bridging the Gap Between Neuromorphic Computing and Machine Learning," Frontiers in Computational Neuroscience, 2021.
[13] A. Fazio, "Advanced Technology and Systems of Cross Point Memory," IEEE Intl. Electron Devices Meeting, 2020.
[14] J.P. Noel et al., "A 35.6 TOPS/W/mm^2 3-Stage Pipelined Computational SRAM With Adjustable Form Factor for Highly Data-Centric Applications," IEEE Solid-State Circuits Letters, vol. 3, 2020.
[15] https://github.com/CEA-LIST/HybroGen