NimbleAI: 3D-Integrated Neuromorphic Vision Sensor-Processor to go where our Eyes can't

Xabier Iturbe
IKERLAN
Arrasate, Spain
xiturbe@ikerlan.es

Gianluca Furano
ESA ESTEC
Noordwijk, The Netherlands
gianluca.furano@esa.int

Didier Keymeulen
NASA JPL
Pasadena CA, USA
didier.keymeulen@jpl.nasa.gov

***Abstract*—AI paves the way for intelligent space agents that make autonomous decisions without the need for Earth-based interaction. Unfortunately, mainstream AI frameworks and algorithms as well as the COTS processors and AI accelerators on which these run are driven by the consumer market, and hence prioritize productivity over efficiency. The *NimbleAI* Horizon Europe project is aimed at redefining the basis of AI-based visual sensing and processing to boost efficiency by leveraging key biological principles in eyes and brains. The project will deliver novel neuromorphic hardware designs and IP to sustain event-driven vision with the following target KPIs: (1) 100x performance per mW gains compared to COTS processors (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. This paper seeks liaison of the project with the space community and stakeholders to get feedback early in the design cycle, and ultimately increase adoption opportunities of NimbleAI technology in next-generation vision payloads.**

I. Introduction

Efficient use of energy, mass and downlink bandwidth are important design drivers for space payloads, especially as small satellites gain momentum in the NewSpace era and ever more ambitious deep space exploration missions are being envisioned. To increase remote sensing and processing capabilities of devices put into space, both business-oriented NewSpace companies and science-driven space agencies are adopting latest generation AI-enabled embedded COTS processors, while eyeing new technology concepts including neuromorphic hardware [2]. Running AI on-board LEO satellites and CubeSats makes it possible to optimize use of limited communication bandwidth for real-time transmission of relevant data to reduce reaction times in Earth observation missions [3]. The need for in-situ processing and on-board sensory perception (especially visual) is even more crucial in deep space exploration missions, where real-time human intervention is not possible because of the long time it takes for data to travel back and forth.

Most COTS processor architectures (e.g., CPU/GPUs) are very inefficient in comparison to biological eyes and brains, which are honed by natural selection. The NimbleAI project [4] leverages key principles of energy-efficient light detection in eyes and visual information processing in brains to create an integral sensing-processing neuromorphic architecture that adopts the biological data economy principle at different levels. Notably, the NimbleAI 3D-integrated architecture implements energy-efficient neuromorphic sensing and processing components along with AI accelerators that support run industry standard AI-based computer vision solutions, thus balancing productivity and efficiency. 3D integration helps overcome the Von Neumann bottleneck in COTS processors, and leverages combined use of the most appropriate process nodes to implement each component in the same chip [6].

This article discusses the main design features and functioning principles of the NimbleAI architecture: (1) sense only significant light changes (visual events) at the optimal spatiotemporal resolution using Dynamic Vision Sensor (DVS) technology; (2) distill sensed visual events to increase information-efficiency; and (3) process selected information-rich events using specialized kernels and AI models.

The article is organized as follows. Section II introduces major challenges in the current state of AI-based computer vision evolution, and Section III outlines the NimbleAI approach to deal with these challenges. Section IV discusses expected benefits of NimbleAI in space applications. and Section V sums up the main takeaways to conclude the paper.

II. Current state of AI evolution and challenges

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

C1.- Complexity of AI models: Accuracy of AI-based computer vision algorithms is commonly opposed to efficiency [7]. Convolutional Neural Networks (CNNs) are typically scaled up to increase accuracy by adding more layers or by enlarging these to process images at a higher resolution. However, to keep workloads manageable by current inefficient processing architectures and fit large CNNs in resource-constrained embedded processors, it is necessary to downscale the input image resolution and shrink the network models, thus sacrificing accuracy.

C2.- Performance and latency: State-of-the-practice computer vision solutions 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

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: www.nimbleai.eu.

completed on the whole frame before the computation of the next layer starts. This results in growing inference delays as AI models 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 [8], 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 [9]. State-of-the-art neuromorphic architectures, on the other hand, implement brain-inspired event-driven Spiking Neural Networks (SNNs) to enormously increase energy-efficiency as they process only changes in their inputs [10]. An important limitation of neuromorphic chips is that the size of SNNs that can run is restricted by the implemented neuron count in silicon. Innatera and Synsense commercial chips implement only 1 k 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 [11], while real-world vision applications typically require 10-20 M neurons.

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 GPUs and NPUs.

III. The NimbleAI Approach

NimbleAI is designing a 3D-integrated sensing-processing architecture that can be customized (at design time and after deployment) to a broad range of computer vision applications and a great variety of deployment environments. The project explores system-level trade-offs in this architecture and pursues specific improvements in selected components where significant energy efficiency margins are foreseen, or novel capabilities are envisioned. As opposed to current event-based vision approaches that are yet limited (e.g., [12]), NimbleAI aims to demonstrate techniques and hardware support to enable seamless combined use of neuromorphic SNNs and industry standard user-trained CNNs and thus achieve powerful yet efficient vision perception.

Fig. 1 shows a conceptual view of the NimbleAI 3D architecture, where the top DVS layer senses light and delivers visual event flows to downstream processing and inference engines in the interior layers. The six layers shown here are only illustrative. Optimal partitioning decisions of components into layers will be made using a 3D design space exploration EDA tool that is being developed in the project, based on [5].

One of the novel system-level concepts in the NimbleAI 3D architecture are *visual pathways* that stream visual event flows from sensor regions occupied by salient image features to processing and inference layers using dedicated Through Silicon Vias (TSVs) [6]. Visual pathways are assigned to Regions of Interest (ROIs) in the sensor in a one-to-one fashion and hence are dynamically created (and destroyed) as new ROIs are detected. They are independently configured (from sensor to processing) at optimal accuracy and latency levels for each ROI. This includes establishing the shortest routes for the event flows and adjusting the optimal working point for the processing and inference engines to serve the workloads with minimal energy use. We posit that visual pathways are an elegant way to answer challenge C4 and harness the increased bandwidth brought about by 3D silicon integration, taking advantage of the irregular distribution of visual information and uneven temporal dynamics in the scene.

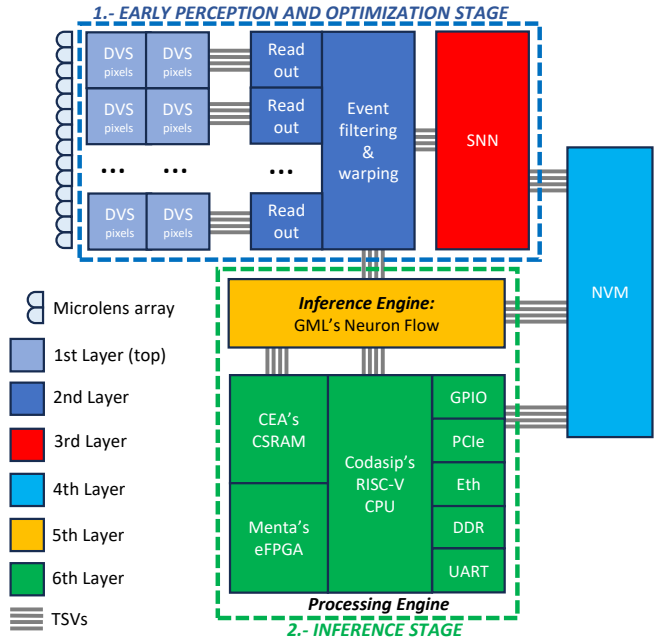

Fig. 1. 3D-stacked conceptual NimbleAI architecture

NimbleAI investigates two novel dynamic vision technologies to improve energy-efficient perception and increase adaptability of visual pathways:

1.- A *digital foveation mechanism* to dynamically allocate sensing resolution in the DVS layer based on the information value brought about by each region to the application. Pixels are dynamically grouped and ungrouped to form macro-pixels with varying resolution levels. This mimics the foveation mechanism in eyes, which allows foveated regions to be seen in greater detail (e.g., salient objects) than peripheral regions (e.g., background). Digital foveation is to be driven by selective attention and optical flow estimations. Hence, several ROIs that match the size, shape and moving dynamics of the recognized and tracked salient objects can be sensed simultaneously. With this, we expect to maximize the amount of meaningful information that can be captured as sensor resolution scales up without unnecessarily increasing the processing workload.

A foveated DVS testchip with expected resolution of 400x400 pixels is being designed and manufactured using X-FAB 180-nm process node.

2.- World's first *event-driven light-field* image sensor is being engineered by coupling a custom-made array of microlenses by Raytrix with a SONY-Prophesee IMX636 DVS [13]. The light-field DVS sensor will run algorithms that 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 edges of objects and that light-field algorithms rely on correlations between neighbour data with lots of redundancy [14], 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 [15]. The project investigates lightweight heuristics to pair DVS events related to incoming light rays based on their temporal activation patterns to circumvent the high computational requirements imposed by frame-based light-field processing. It is expected that depth information will open new opportunities for ROI detection (e.g., nearby objects) and improve end-to-end perception accuracy.

Hence, NimbleAI is studying techniques to capture and optimally represent the spatiotemporal 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. For instance, the 3D information encoded in the event flows is expected to expose new forms of sparsity when applying the right pre-processing kernels (e.g., optical flow based event warping [16]). Therefore, the expectation is that by investing some computing power and energy to gain some situational awareness and shape visual event flows early, a major reduction of the subsequent workload will be achieved, saving lots of energy by doing that.

As shown in Fig. 1, the project envisions a two-stage perception approach, where the two stages reinforce each other to perform more efficiently as the deployment environments become more familiar and visual stimuli are better understood.

1.- An always-on *early perception and optimization stage* implements selective attention and optical flow algorithms to detect, delimit and track 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) for efficient end-to-end region inference. This stage is inspired by unconscious visual processing and neural signalling in biological systems, and hence is largely invisible to the application yet adjustable through user-driven directives.

NimbleAI explores low-energy and low-latency advantages of SNNs, especially when applied to DVS events, to power autonomous functioning of the early perception and optimization stage. Furthermore, SNNs are particularly well suited to online training as their event-based learning rules typically use only information local to the synapse [17]. This property will be conveniently exploited to continuously improve on dynamically selecting ROIs in the early perception stage, using inference feedback from subsequent AI models to drive the SNN online training process. Hence, part of the energy consumed to complete end-to-end inference also serves the purpose of improving the overall energy use.

2.- An on-demand *inference stage* implements pre- and post-processing kernels on the downstream processing engine (i.e., Codasip's RISC-V CPU, Menta's eFPGA fabric, and CEA's CSRAM in-memory computing memory blocks [18]) and runs user-trained CNNs on the inference engine (i.e., event-driven dataflow GML's NeuronFlow [19]). As it occurs with SNNs, the type of events that are processed by event-driven dataflow architectures such as NeuronFlow correspond with visual events delivered by DVS sensors, thus maximizing end-to-end efficiency along visual pathways. Recent research has shown that 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 [20]. To support visual pathways and optimally benefit from DVS foveation, the inference engine can run multiple CNNs with varying topologies simultaneously.

To deal with challenge C3, NimbleAI explores the novel concept of *Virtual Neural Networks (VNNs)* to enable run large and accurate event-driven CNNs and SNNs in only 50 mm^2 chips. This concept will be supported by dedicated TSVs and stacked layers of Non Volatile Memory (NVM), which will be 3D-architected to create a high-bandwidth and high-density memory hierarchy for quickly swapping active and non-active neurons and layers of running CNNs in the inference engine and SNNs in the early perception stage.

Note that event- and region-based sensing and processing described above helps limit the complexity and energy-consumption of AI models, and thus deal with challenges C1 and C2. AI models and processing kernels 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 to process significant changes in their neuron states and visual inputs. As opposed to state-of-the-practice, where input images are typically downscaled to keep workloads manageable, NimbleAI will process selected full-resolution ROIs for better accuracy. Also, as opposed to state-of-the-practice where more complex/accurate AI models translate directly into more computing and energy consumption, in NimbleAI model complexity to workload translation will be dynamically adjusted through runtime optimization mechanisms that control event generation and processing 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 AI models.

IV. Expected Benefits in Space Applications

NimbleAI expects to deliver technology advantages to build miniaturized, powerful and energy-efficient vision payloads, tackling the challenges introduced in section II. Some of these advantages are linked to event-based vision, which is actively being explored by ESA, NASA and other space agencies [21].

A1.- Data-efficiency (vs C1, C2, C3): NimbleAI is aimed at capturing and processing minimal amounts of data with high-value information to support efficiently scale up AI models and sensor resolution while optimising use of downlink bandwidth.

A2.- Low-latency (vs C2): NimbleAI allows for capturing and processing in real-time high-speed events that cannot be captured with traditional space borne cameras. This is very interesting for 3D mapping of planetary surfaces, monitoring phenomena with fast temporal dynamics, such as explosive eruptions, as well as sporadic events that occur over a very short span of time, such as meteoroids. Low-latency visual inference results are also extremely important in landing, obstacle avoidance, autonomous rendezvous and docking manoeuvres to precisely estimate time-to-contact [22].

A3.- Energy-efficiency (vs C3): NimbleAI is aimed at implementing different runtime optimization strategies to dynamically trade off performance and accuracy in different energy availability situations. This is relevant for instance in the case of LEO satellites, whose (solar) energy availability varies with the orbit inclination [23].

A4.- 3D-Integrated sensing-processing (vs C4): In addition to amplifying the advantages above, 3D integration will enable significant mass reduction in space payloads as a result of circumventing board-level component integration and routing. Payload miniaturization is extremely important as small satellites and CubeSats revolutionize the space industry.

A5.- Superior 3D visual sensing: NimbleAI augments the inherent high dynamic range of DVS technology, allowing to run processing kernels tailored to the specific ranges of lighting in each ROI. This is relevant specially in Earth observation missions [24] and rendezvous maneuvres, where lighting conditions change very quickly based on relative position of spacecrafts with regard to a few high intensity light sources. The NimbleAI light-field DVS sensor is expected to enable passive, high dynamic range, instantaneous, and energy-efficient 3D visual sensing that is largely not affected by lighting (and weather conditions), thus overcoming major LiDAR weaknesses [25].

A6.- After-deployment adaptability: The eFPGA fabric in the NimbleAI architecture provides hardware flexibility to support in-flight upgrades of AI models and processing kernels. This helps keep up pace with rapidly evolving AI-based computer vision algorithms.

V. Takeaways

NimbleAI takes inspiration from ultra energy-efficient eye-brain systems. The project expects to achieve 100x energy-efficiency improvement and 50x latency reduction w.r.t. CPU/GPUs processing frame-based video. To achieve this, NimbleAI enables event-driven visual inference, where sensing and processing are dynamically adjusted to operate jointly at the optimal temporal and data resolution levels.

The project will deliver a functional prototype of the 3D-integrated NimbleAI sensing-processing architecture along with the corresponding programming tools and OS drivers to enable users run their AI models on it. The prototype will be flexible to accommodate user IP and will combine commercial neuromorphic chips and NimbleAI testchips (e.g., foveated DVS sensor). Please reach out to test combined use of your vision pipelines and NimbleAI technology in this prototype.

References

[1] TechEdSat-13: The First Flight of a Neuromorphic Processor, https://ntrs.nasa.gov/citations/20220005780.

[2] D. Keymeulen et al., "High-Performance Embedded System-on-a-Chip for Space Imaging Spectrometer," IEEE Aerospace Conference, 2023.

[3] X. Iturbe et al., "NimbleAI: Towards Neuromorphic Sensing-Processing 3D-integrated Chips," ACM/IEEE Conf. on Design, Automation and Test in Europe, 2023.

[4] https://www.monozukuri.eu/genio-3d

[5] G. Van der Plas and E. Beyne, "Design and Technology Solutions for 3D Integrated High Performance Systems," Symp. on VLSI Circuits, 2021.

[6] M. Tan et al., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Intl. Conf. on Machine Learning, 2020.

[7] J. Shalf, "The future of computing beyond Moore's Law," Philosophical Transactions of the Royal Society, 2020.

[8] N.P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," Annual Intl. Symp. on Computer Architecture, 2017.

[9] C.D. Schuman et al., "Opportunities for Neuromorphic Computing Algorithms and Applications," Nature Computational Science, 2022.

[10] P. Merolla et al., "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," SCIENCE, vol. 345, no. 6197, 2014.

[11] J. Hagenaars et al., "Self-Supervised Learning of Event-Based Optical Flow with Spiking Neural Networks," Intl. Conf. on Neural Information Processing Systems, 2021.

[12] https://www.sony-semicon.com/en/news/2021/2021090901.html

[13] F. Bhatti and T. Greiner, "Design of an FPGA Hardware Optimizing the Performance and Power Consumption of a Plenoptic Camera Depth Estimation Algorithm," MDPI Algorithms, vol. 14, no. 7, 2021.

[14] C. Kim et al., "Scene Reconstruction from high Spatio-Angular Resolution Light Fields," ACM Transactions on Graphics, vol. 3, no. 4, 2013.

[15] S. Shiba et al., "Secrets of Event-Based Optical Flow," European Conf. on Computer Vision, 2022.

[16] J.L. Lobo et al., "Spiking Neural Networks and Online Learning: An overview and perspectives," Neural Networks, vol. 121, 2020.

[17] 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.

[18] O. Moreira et al., "NeuronFlow: a Neuromorphic Processor Architecture for Live AI Applications," Conf. on Design, Automation and Test in Europe, 2020.

[19] L. Deng et al., "Understanding and Bridging the Gap Between Neuromorphic Computing and Machine Learning," Frontiers in Computational Neuroscience, 2021.

[20] https://www.esa.int/gsp/ACT/projects/event_camera

[21] O. Sikorski et al., "Event-based spacecraft landing using time-to-contact," IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2021.

[22] K. B. Chin et al., "Energy Storage Technologies for Small Satellite Applications," Proc. of the IEEE, vol. 106, no. 3, 2018.

[23] C. Cullingworth and J.P. Muller, "Contemporaneous Monitoring of the Whole Dynamic Earth System from Space," MDPI Remote Sensing, vol. 13, no. 5, 2021.

[24] Y. Li and J. Ibanez-Guzman, "Lidar for Autonomous Driving: The Principles, Challenges, and Trends for Automotive Lidar and Perception Systems," IEEE Signal Processing Magazine, vol. 37, no. 4, 2020.