SENECA: Flexible and Scalable Neuromorphic Processor

Guangzhi Tang, Kanishkan Vadivel, Yingfu Xu, Kevin Shidqi, Paul Detterer, Stefano Traferro,
Mario Konijnenburg, Manolis Sifalakis, Ger-Jan van Schaik, Amirreza Yousefzadeh
imec Netherlands, Eindhoven, The Netherlands

SENECA

SENECA (**S**calable **E**nergy-efficient **NE**uromorphic **C**omputer **A**rchitecture) [1] is a **RISC-V-based** digital neuromorphic processor targeting extreme edge and near-sensor applications:

- ***Energy-Efficient*** event-driven computation exploit activation sparsity in edge neural networks.
- ***Low-Latency*** parallel processing via single-core SIMD and multi-core asynchronous execution.
- ***On-device Adaptation*** enables privacy-aware learning and increases robustness in deployment.

Comparison with State-of-the-art Neuromorphic Designs

SENECA improves state-of-the-art neuromorphic designs in following aspects:

- ***Flexible*** programmable neural models, synaptic models and learning algorithms.
- ***Area-Efficient*** 3-level memory hierarchy allows novel embedded memory technologies.
- ***Multi-Precision*** graded spikes, weights, and neural states in 3 data types: int4, int8, BF16.
- ***Core-to-Core Asynchronous Parallelism*** without global synchronization overhead.
- ***End-to-End Application Deployment*** with pre/post and main network processing in one go.

Architecture	Energy/SOp *(pJ)*	Area *(mm2/Core)*	Memory *(Mb)*/Core	Flexibility
SENECA	**2.8**	**0.47**	**2.0**	**High**
Loihi2	N/A	0.21	1.5	Medium
SpiNNaker2	10	1.09	1.0	High
Tianjic	1.54	0.09	0.17	Low
ReckOn	5.3	0.45	1.1	Fixed

Table 1. Neuromorphic state-of-the-art comparison

Flexible and Scalable Neuromorphic Architecture

- *Novel hierarchical control system* consists of RISC-V and Loop Buffer guaranties flexibility and efficiency at the same time.
- *Neuron Processing Element array (NPE)* accelerates neural network computation.
- *Network-on-Chip (NOC)* with multicasting, compression mechanism, source-based routing.
- *Customizable and scalable digital IP* on number of cores, NPE array size, memory size, targeting a wide range of applications.

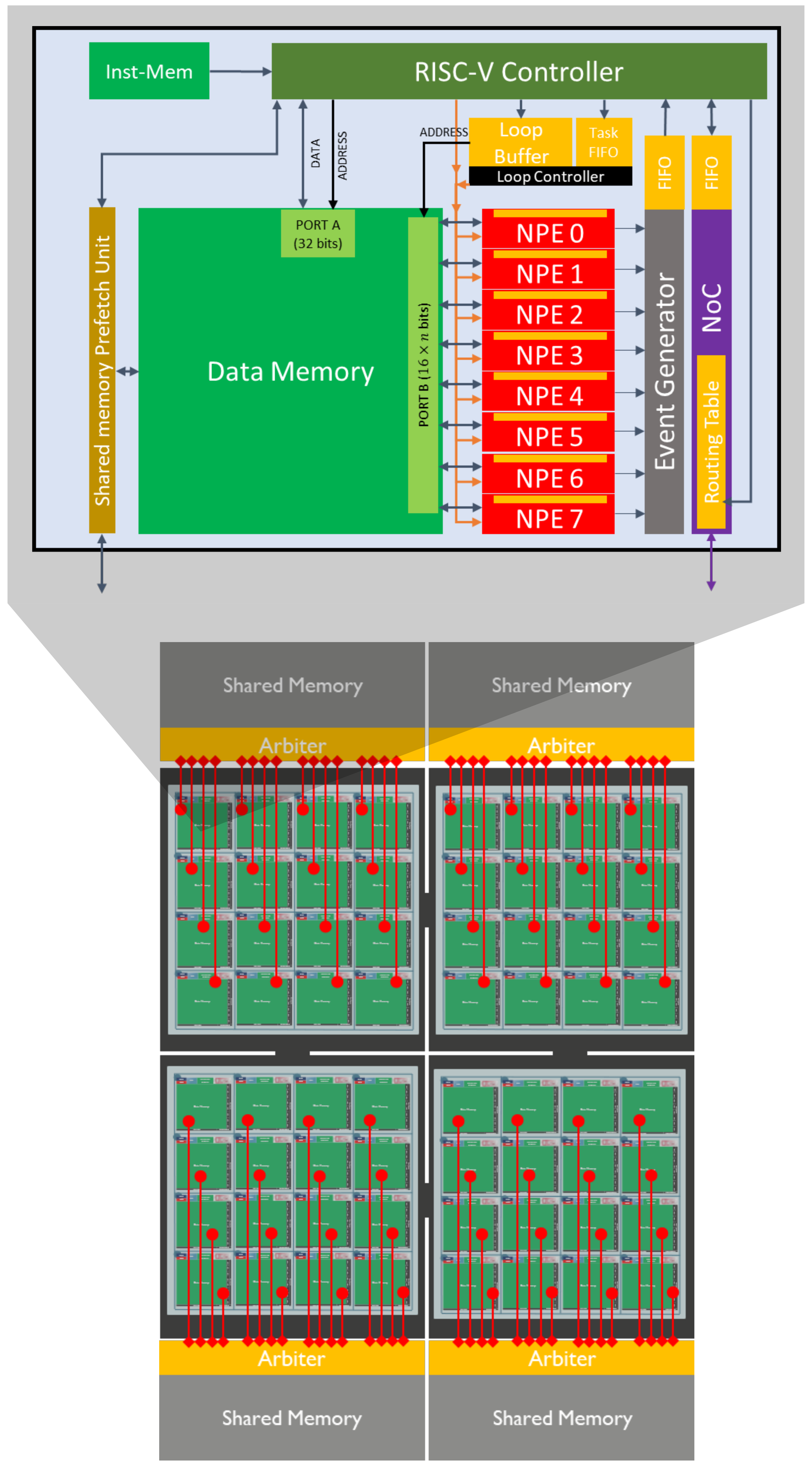

Figure 1. Single-core (top) and 64-core (bottom) SENECA architecture

Sparsity-aware Event-driven Processing

- *Event-driven processing* efficiently exploits fine-grained activation sparsity.
- *Spike grouping* simultaneously process multiple events to reduce costly memory access.
- *Low-precision activation and weight* support enlarge search space for efficient network design.

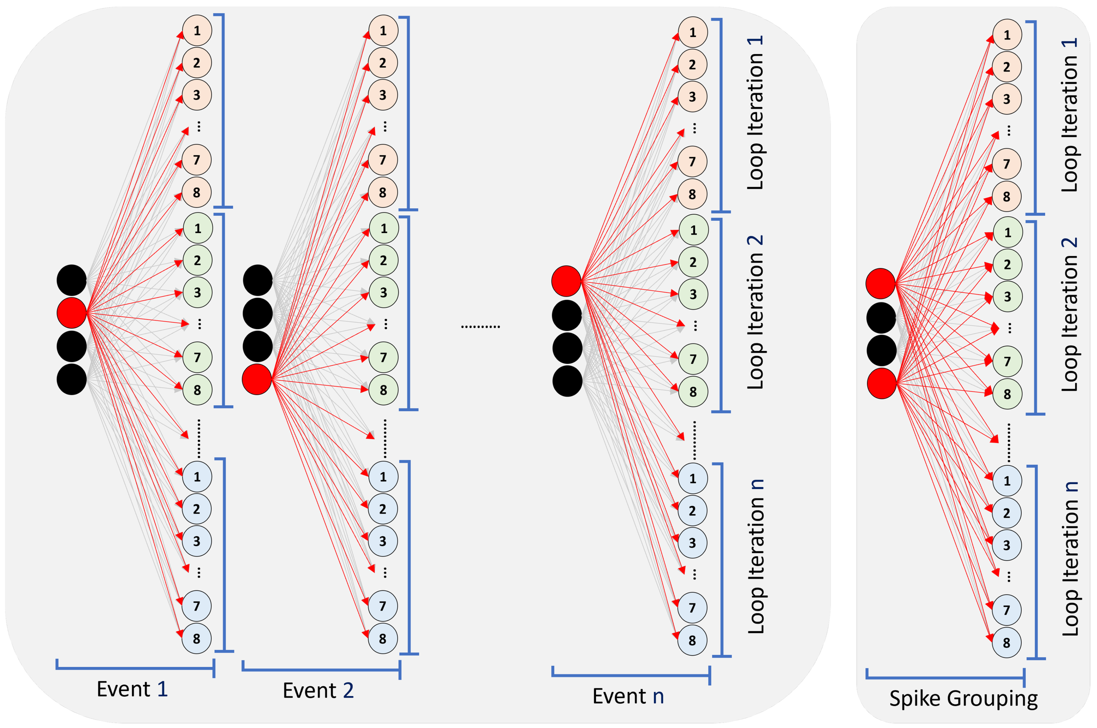

Figure 2. Event-driven processing and spike grouping optimization

	Binary Spike +low-precision +spike-grouping	Binary Spike	Graded Spike
Latency	**1x**	3.1x	4.0x
Energy/SOp (pJ)	**1x (2.7)**	4.9x (13.3)	5.3x (14.2)

Table 2. Event-driven processing optimization

Event-driven Convolutional Neural Layer

- *Memory-efficient processing* via weight and state reuse for high-resolution event cameras.
- *Minimum overhead* on pre- and post-processing benefit from our novel hierarchical control system.

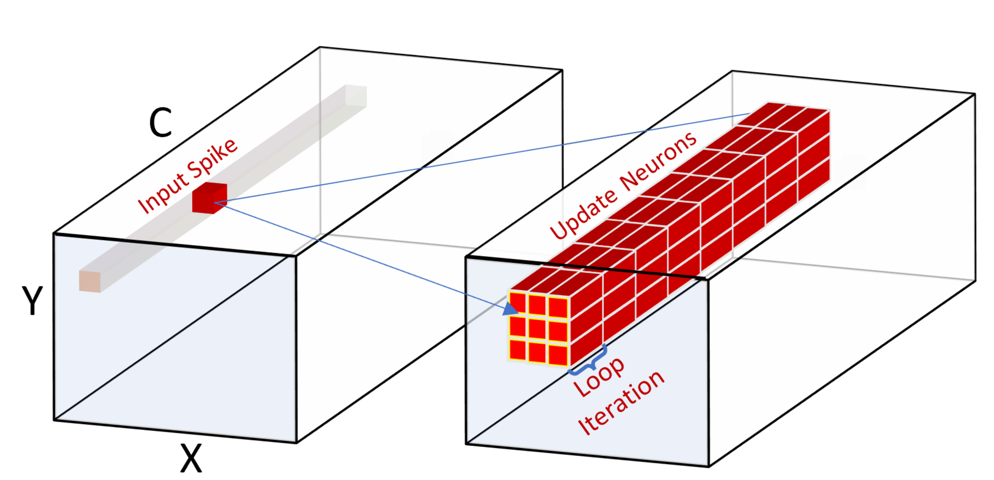

Figure 3. Event-driven convolutional layer processing

On-device Adaptation and Learning

- *Bio-inspired learning* with limited overhead [2] on network inference by avoid using backpropagation.
- *On-device adaptation* for senser aging, domain shifting, new class adding, via few-shot learning and continual learning.

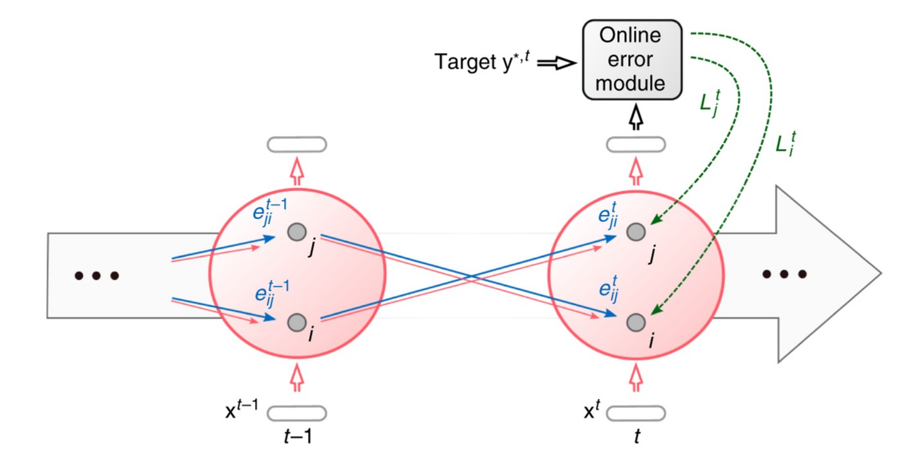

Figure 4. Online recurrent learning of e-prop (from [Bellec et al., 2020])

Collaboration and Contact

SENECA platform is accessible for academic research and welcome industrial collaboration, please contact us!

Guangzhi Tang (presenter), Researcher at imec Netherlands, guangzhi.tang@imec.nl (Holst Center, HTC 31, Eindhoven, The Netherlands)

Publications

[1] SENECA: Building a fully digital neuromorphic processor, design trade-offs and challenges. To be appeared in *Frontiers in Neuroscience*, 2023.

[2] Open the box of digital neuromorphic processor: Towards effective algorithm-hardware co-design. *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023.

Acknowledgement

This technology is partially funded and initiated by the Netherlands and European Union's Horizon 2020 research and innovation projects **TEMPO** (ECSEL Joint Undertaking under grant agreement No 826655), **ANDANTE** (ECSEL Joint Undertaking under grant agreement No 876925), **DAIS** (Key Digital Technologies Joint Undertaking under grant agreement No 101007273) and **REBECCA** (Key Digital Technologies Joint Undertaking under grant agreement No 101097224).