SENeCA: Next generation Neuromorphic Processor for EdgeAI

Amirreza Yousefzadeh, Gert-Jan van Schaik, Mohammad Tahghighi, Paul Detterer, Stefano Traferro, Manolis Sifalakis, Martijn Hijdra, Jan Stuijt, Matthias Hartmann, Francky Catthoor, Ilja Ocket, Mario Konijnenburg

SENeCA

- SENeCA is

  a **RISC-V**-based digital neuromorphic processor targeting extreme edge applications by accelerating Spiking Neural Networks inside or near sensors and small devices where ultra-low power and adaptivity are required

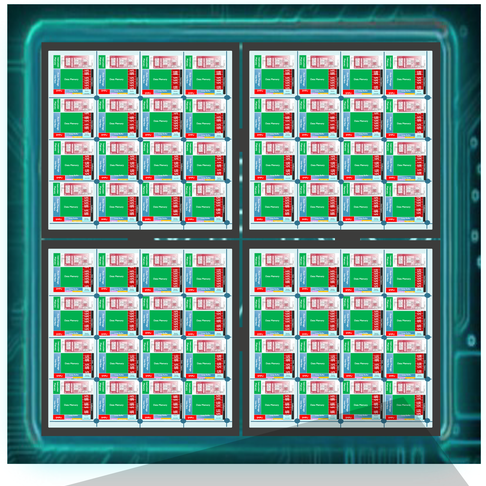
- SENeCA inherits fundamental properties from the biological brain:
  - Spatio-temporal sparsity exploitation
  - Parallel processing
  - Infinite scalability
  - Low-precision parameters
  - Asynchronous - Non-deterministic execution
  - Adaptation and fault-tolerance

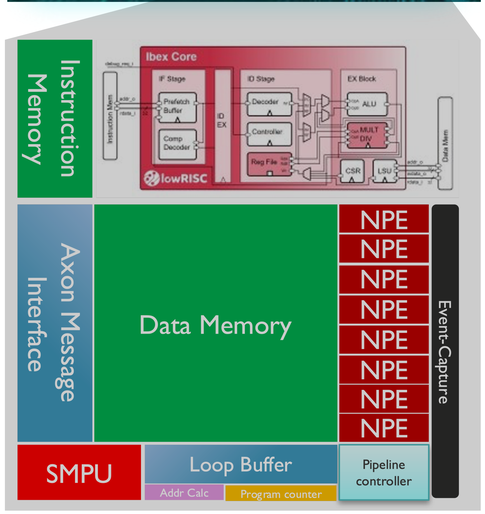
Main Features

- SENeCA improves current state of the art
  - **Flexibility**: fully programmable neuron models and learning/adaptivity algorithms;
  - **Area efficiency**: 3-level memory hierarchy allows novel embedded memory technologies;
  - **Advanced learning mechanisms and optimization algorithms**: neural operations in three data types: int4, int8 and BrainFloat16;
  - **Efficient event communication**: Network-on-Chip with multicasting, a compression mechanism, and source-based routing.
  - **End2End application deployment**: Pre/Post and the main processing

Architecture

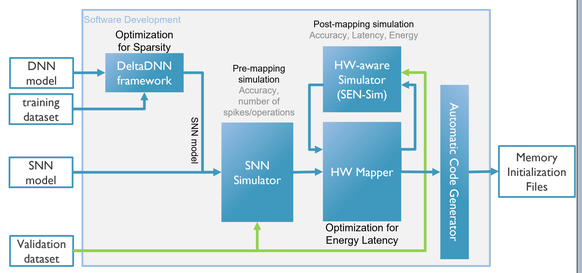
- Interconnect of Neuron Cluster Cores, with RISC-V-based instruction set,
- Optimized Neuromorphic Co-Processor,
- Event-based communication infrastructure.
- Customizable digital IP for different applications
  - number of cores
  - Neural Processing Elements (NPEs) per core
  - optional use of off-chip memory



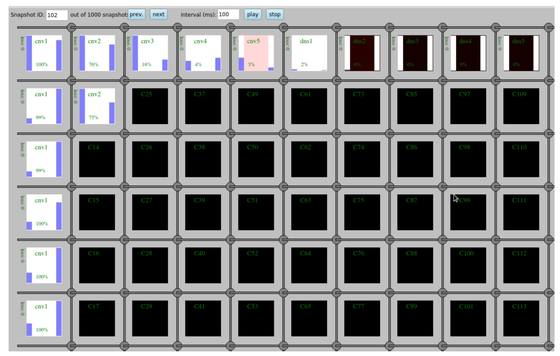


SDK

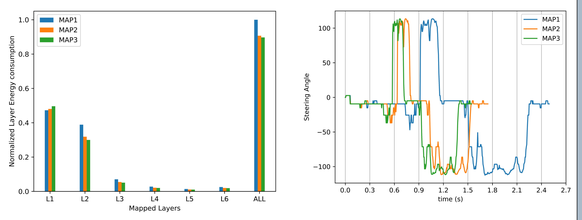
- SENeCA platform SDK provides:
  - SNN optimizer to improve sparsity
  - Pre-mapping SNN simulation
  - Post-Mapping (hardware-aware) simulation
  - Close-loop mapping optimization
  - Automatic code generation for RISC-V controllers



SDK Architecture



SDK GUI



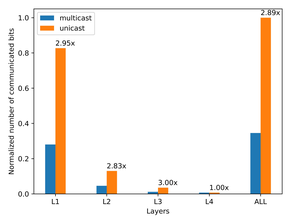
Mapping effect on Energy/Latency

Synthesis Results

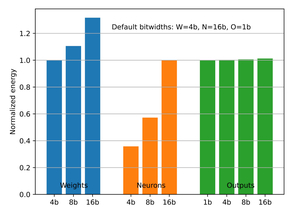
- Early synthesis results in GF22nm: 0.55mm2/core (80% SRAM) for
  - 8 physical neurons
  - 32K virtual neurons
  - 256K parameters (4b)

Early Benchmarking results

- 45% energy reduction by using level-0 memory hierarchy (16-word Register File) from maximum data-reuse
- 60% less communication bits, by using event-compression mechanism and multicasting NoC



Multicasting NoC effect on communicated bits



Parameter quantization vs Energy

Target applications

- Edge Audio/video processing
  - Denoising in portable Ultra-Sound imaging devices
  - Smart office buildings (Sound-based Emotion Recognition)
- Asynchronous sensor fusion
- Event-based data processing (e.g., a processor for Dynamic Vision Sensors)

The platform is accessible for academic research, please contact us!

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) and **DAIS** (Key Digital Technologies Joint Undertaking under grant agreement No 101007273).