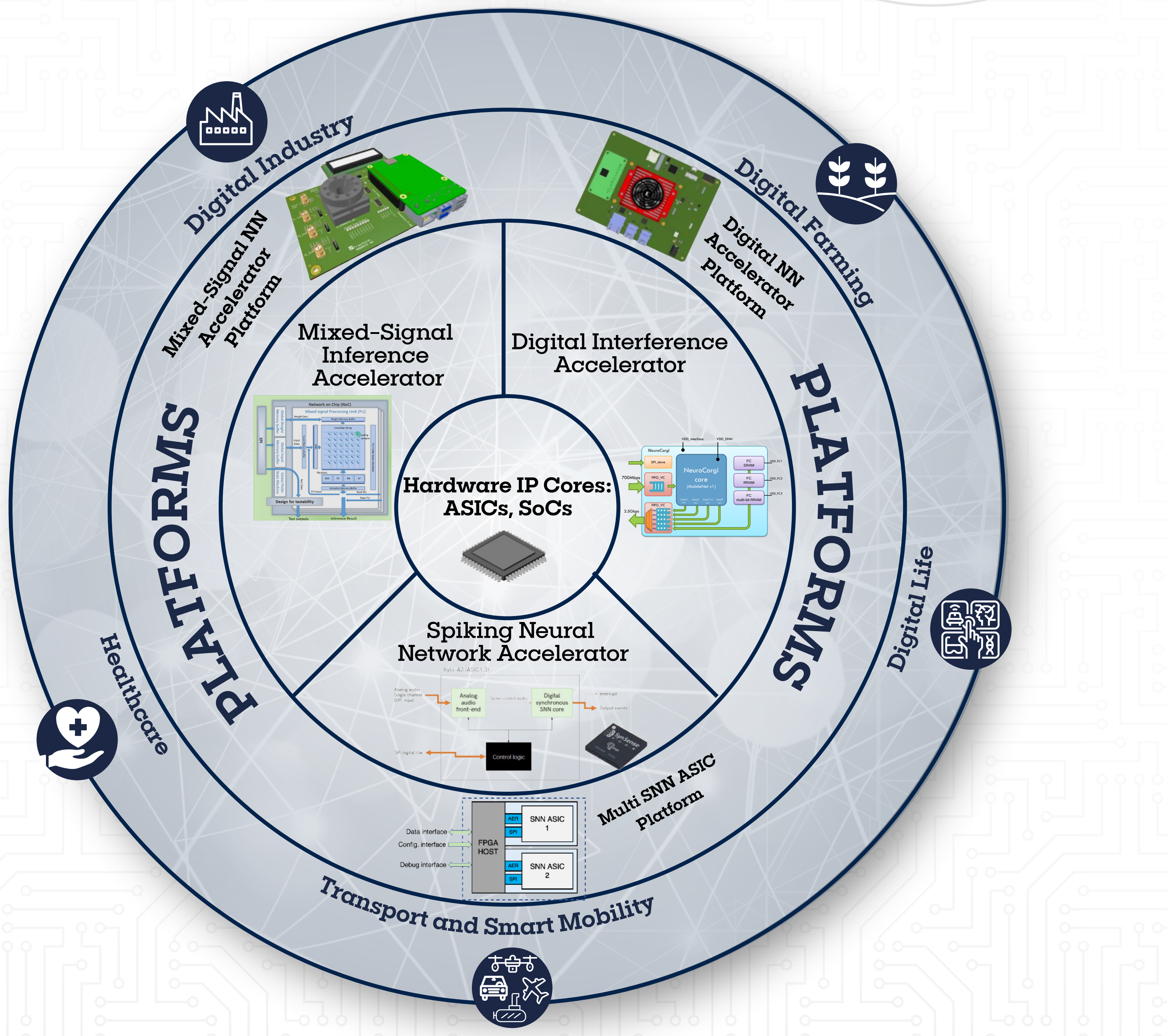
ANDANTE

HW Design of Edge AI processors implementing Spiking and Artificial Neural Networks using different embedded Non-Volatile Memory technologies.

Introduction

ANDANTE aims to design and implement efficient hardware AI accelerators to perform AI functions like classification and segmentation based on analog, digital and spiking neural networks for Edge applications. Fourteen different designs (ASICs, SoCs and FPGAs) have been implemented, some of which including new embedded non-Volatile memories (FeFET, OXRAM, PCM) and in memory computing IPs in some of the designs to optimize the energy efficiency.

These designs are the basis of fourteen inference deployments in five different industrial domains, which are used to validate performances of these circuits at the Edge.

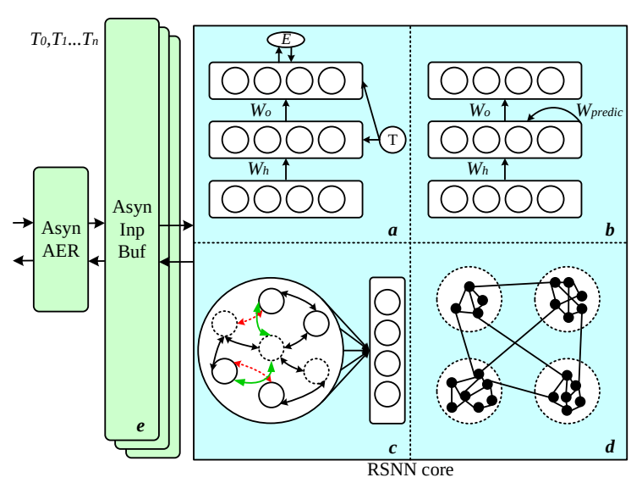

Digital Industry
Digital Farming
Digital Life
Transport and Smart Mobility
Healthcare
PLATFORMS
PLATFORMS
Mixed-Signal NN Accelerator Platform
Digital NN Accelerator Platform
Multi SNN ASIC Platform
Mixed-Signal Inference Accelerator
Digital Interference Accelerator
Hardware IP Cores: ASICs, SoCs
Spiking Neural Network Accelerator

SNN Core with On-Chip Learning

ASIC 1.2 Always-on SNN Core

- **Target**: Energy and area efficient SNN processor with on-chip learning
- **Technology**: ST/P28
- **Memory type**: 160kB SRAM
- **NN Type**: Spiking Neural Network(recurrent/feedforward) with max.512 neurons running on real time or accelerated
- **Status**: Under design, **TapeOut** Q4/2023
- **Feartures: ab.** Supervised/sel-supervised on-chip learning with local plasticity c. Dynamic sparse training d. Small-world network mapping e. Mixed-timing design
- **Key features compared to SOTA SNN chips**:
  - √ More hardware friendly on-chip learning
  - √ Sparse weight matrix radically reduces the on-chip memory footprint
  - √ Mixed-timing design combines advantages of both asyn. and syn. circuits
  - √ NVM PCM to store the parameters
  - √ CIM macro to replace time-multiplex SRAM
- **Silicon**: Q1/Q2 2024; Area: ~3 mm²
- **Expected power consumption**: Less than 100µw
- **Use case**: Low-dimensional signal inference and learning, such as auditory signals, bio-signals and simple vision task

Block Diagram of SNN Core with On-Chip Learning

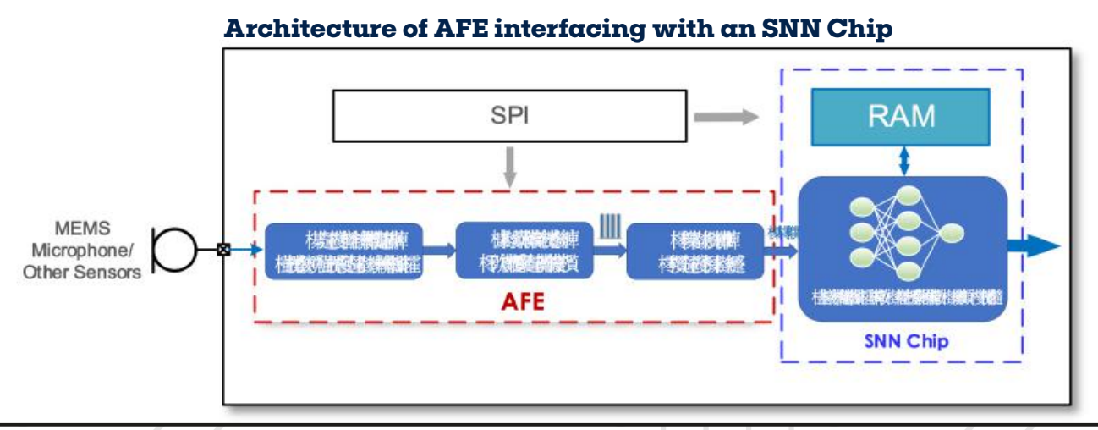


Audio Front-end

ASIC 1.3 Audio Processing

- **Target**: Low -power analog audio front-end tailored for human speech applications. It performs analog filtering via a set of tuneable bandpass filters for its processing by a Spiking Neural Network. not related to machine vision
- **Technology**: 40 nm TSMC
- **Input bandwidth: 100 – 20 kHz**
- **Expected power consumption**: < 300 uW
- **Status**: Silicon available
- **Use case**: 3.2 Under water signal classification, 5.1 Consumer Auditory processing

Architecture of AFE interfacing with an SNN Chip

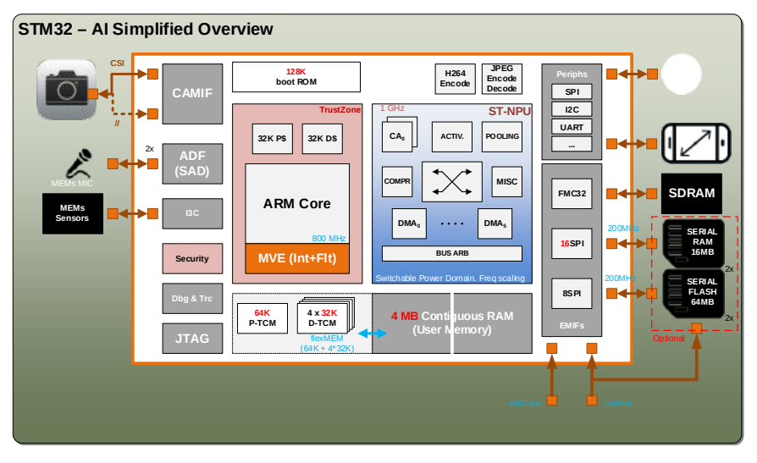


Digital MCU with AI

SoC 1.1 ST32 –AI MCU

- **Target**: SoC combining a Microcontroller STM32 microcontroller with AI acceleration via a neural processing unit for consumer applications
- **Technology**: 16 nm FinFET TSCM / SRAM
- **Memory**: SRAM, 4.2 Mbyte
- **NN type**: ANN/CNN/Yolo V3
- **Expected efficiency**: 3.3 TOPS/W
- **Expected power consumption**: < 300 mW
- **Status**: Silicon available
- **Measured Performances**:
  - 314 fps
  - 1 to 2 order of magnitude > SW solution on STM32H7 (Arm Cortex A7)
- **Use case**: consumer applications

Block diagram of STM32 AI SoC

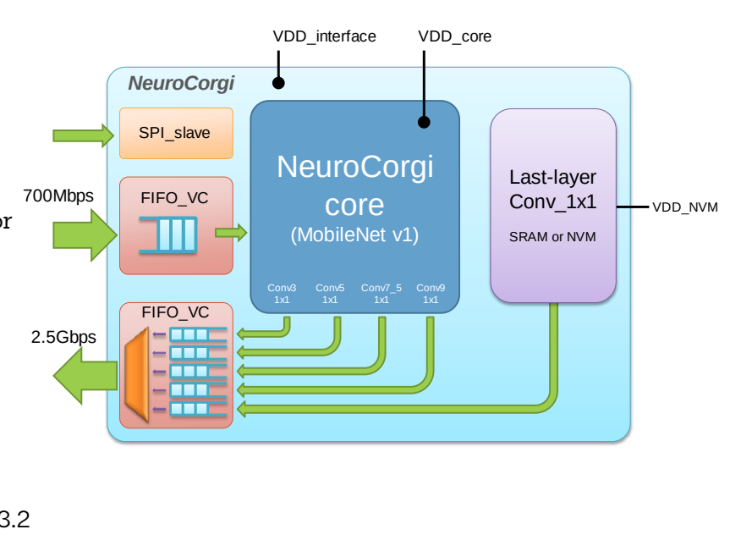


Digital ANN

ASIC 2.1 NeuroCorgi

- **Target**: Feature extractor circuit to address image classification, segmentation and detection for ANN applications minimizing the energy required per inference while having an extremely low latency.
- **Technology**: 22 nm FDSOI GF
- **Memory**: SRAM, ~600 kbyte / OxRAM,
- **NN type**: ANN/CNN/MobilNet V1
- **Input throughput**: images 1280 pixels (24 bits RGB pixels), 1280x720 @30 FPS or 1280x360 @60 FPS or 640x480 @90 FPS
- **Inference Latency**: < 10 ms
- **Expected efficiency**: > 10 TOPS/W
- **Expected power consumption**: < 100 mW
- **Status**: Under testing
- **Silicon**: June 26th, 2023
- **Use case**: 2.1: Autonomous Weeding System, 2.2: Tomate Pest and disease forecast, 3.1: Drones/USV, 3.2 Under water signal classification; 3.3: 3D object detection and classification of road users based on Lidar and camera, 3.4: Robust autonomous landing

NeuroCorgi Architecture

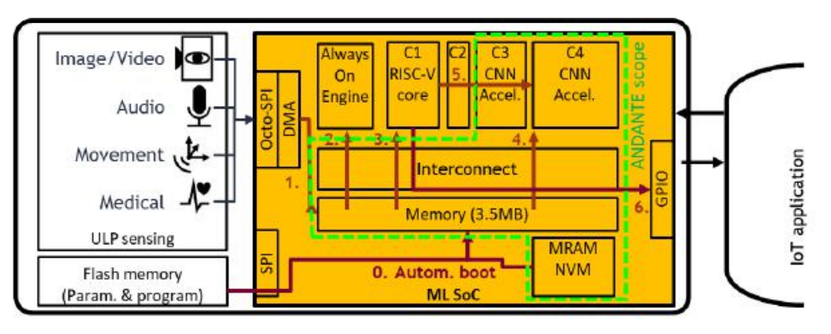


Digital CNN

SoC 2.1 Visage 2

- **Target**: Neural Compute Engine (NCE) targeting NN acceleration for smart vision applications in digital life domain
- **Technology**: 22 nm FDSOI GF
- **Memory**:3.5 MB SRAM, 0.5MB MRAM
- **NN type**: ANN/diverse classes
- **Input throughput**: OctoSPI @ 1600 Mbps DCMI @ 500 Mbps
- **Expected efficiency**: 10 TOPS/W
- **Expected power consumption**: 10 mW to 10 mW
- **Status**: Under design, Fab-in Q1 2024
- **Silicon**: Q2, 2024
- **Use case**: 5.2 vision–based human computer interaction applications

Architecture of End-to-end ML inference SoC with the NCE for ML acceleration

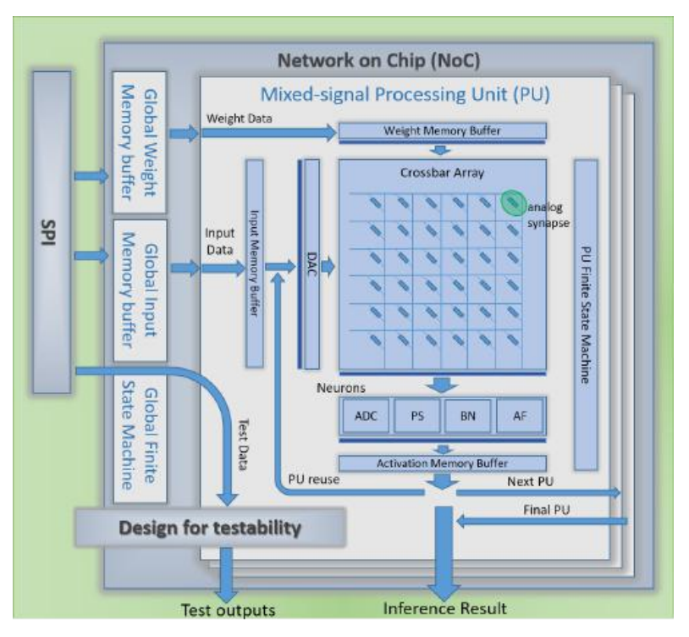


IIS&EMFT: Mixed Signal ANN

ASIC 3.1 Adelia 22 gen2

- **Target**: Scalable and configurable mixed-signal inference accelerator with a multi-core architecture using analog in-memory computing for voice activity detection (VAD).
- **Technology**: 22 nm FDSOI GF
- **Memory**: SRAM , ~600 kB
- **NN type**: ANN
- **Input throughput**: Audio features up to 64x13x20x 8 bit/s
- **Inference Latency**: < 10 ms
- **Accuracy**: 82% min
- **Expected efficiency**: ~5 TOPS/W (estimated for 8b OP)
- **Expected power dissipation**: 1mW
- **Status** : Silicon under test.
- **Use cases**: 5.1d Voice activity detection (VAD).

Architecture of the ADELIA 22 Gen2 ASIC

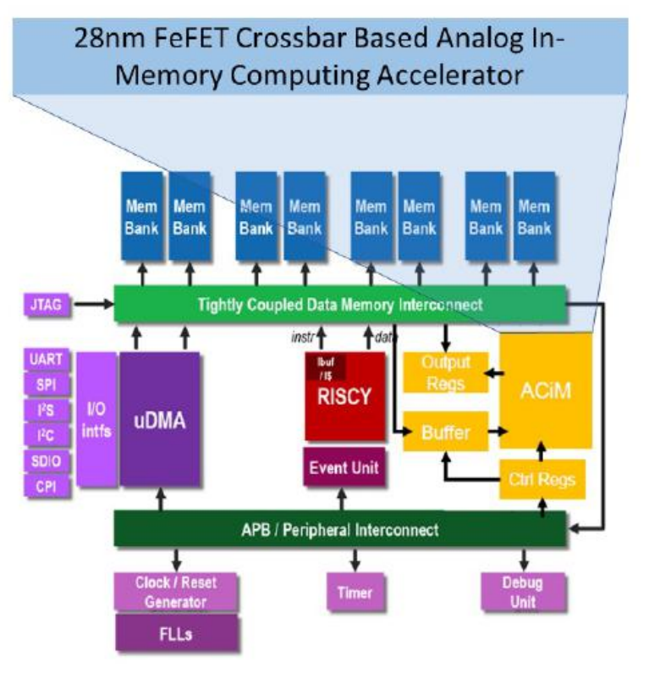


IPMS: Mixed Signal ANN

ASIC 3.1b IMC

- **Target**: Flexible SoC for convolutional neural networks integrating multiply-accumulate (MAC) accelerators using FeFETs with a RISC-V microcontroller for person detection and classification.
- **Technology**: 28 nm SLPe GF
- **Memory**: FeFET
- **NN type**: ANN/CNN/Yole-V3-Tiny
- **Throughput**: 20 inference/s typ
- **Inference latency**: 10 ms
- **Accuracy**: 61.3% tmAP
- **Expected efficiency**: 20 TOPS/W
- **Expected power dissipation**: 10 mW
- **Status**: Fab-out July 2023.
- **Use case**: 1.1 People counting and indoor positioning

Architecture IMC SoC

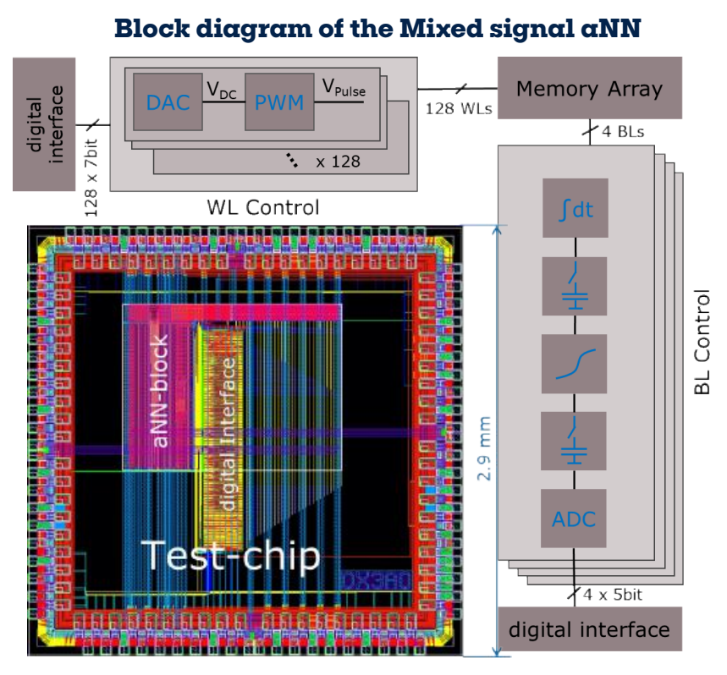


Analog Neural Network (aNN)

ASIC 3.2 aNN IMC

- **Target**: Analogue neuronal network (aNN) for tinyML applications. To be evaluated in the context color recognition.
- **Technology**: 28 nm HPC+ TSMC
- **Memory**: RRAM , < 20 kbytes
- **NN type**: aNN
- **Input throughput**: 128 x 7 bit
- **Inference Latency**: 5 ms max
- **Accuracy**: 85% min
- **Expected efficiency**: < 1 GOPS/W
- **Expected power dissipation**: 5 mW max
- **Status**: Silicon available and validated
- **Use cases**: UC1.2 color recognition

Block diagram of the Mixed signal aNN

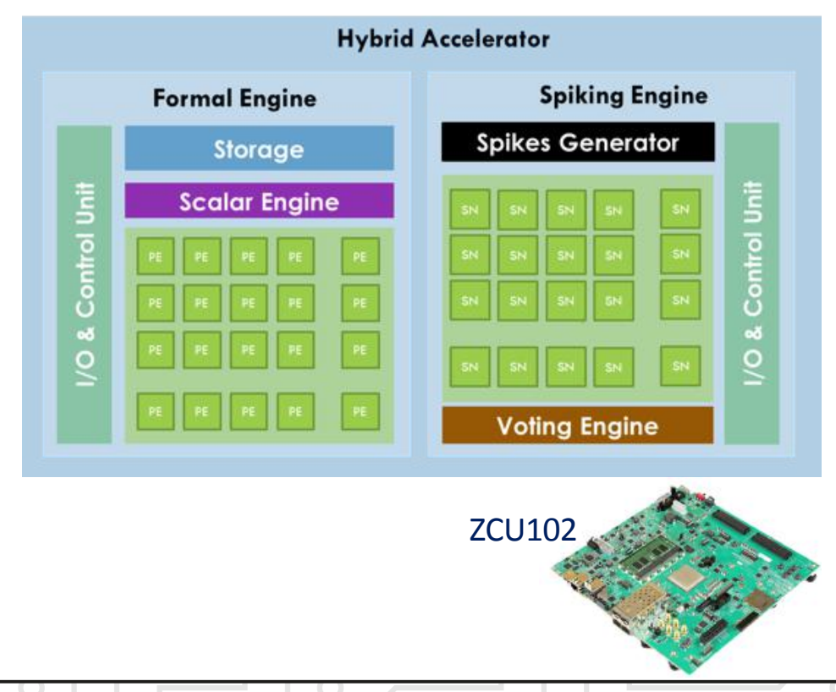


Detection, Classification and Segmentation of High Altitude Images

FPGA-1 : Hybrid Accelerator

- **Target**: Detection, classification and segmentation of high altitude images using either ANN, SNN or a hybrid technology
- **FPGA**: Zynq UltraScale+ MPSoC ZCU102
- **Memory**: up to 10MB
- **Cores**: >1000
- **NN type**: ANN, SNN (IF)
- **Accuracy**: depends on the algorithm
- **Status**: FPGA under validation
- **Expected power consumption**: <20W Use cases: 3.1 Drones/USV

Architecture

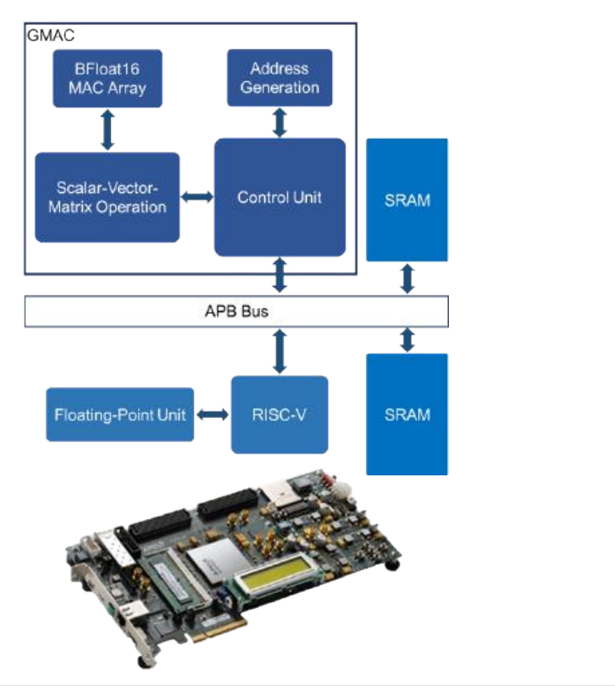


Online Learning Accelerator

FPGA-2 : GMAC

- **Target**: Highly configurable general 16-bit floating-point online learning hardware accelerator (GMAC) for Recurrent Neural Network, Spiking RNN and Multilayer Perceptron tasks. Prototype for integration in next-generation SpiNNaker2 system.
- **FPGA**: Virtex-7 / VC707
- **Memory**: 128KB
- **NN type**: RNN/SRNN/MLP
- **Accuracy**: Depends on Algorithm
- **Speedup**:
  - **27x** for pointwise Matrix operations
  - **44x** for Vector-Matrix-Multiplication
  - **81x** for Transpose Vector-Matrix-Multiplication
- **Status**: FPGA under validation
- **Use cases**: 1.1 People counting and indoor positioning

Architecture of the GMAC SOC

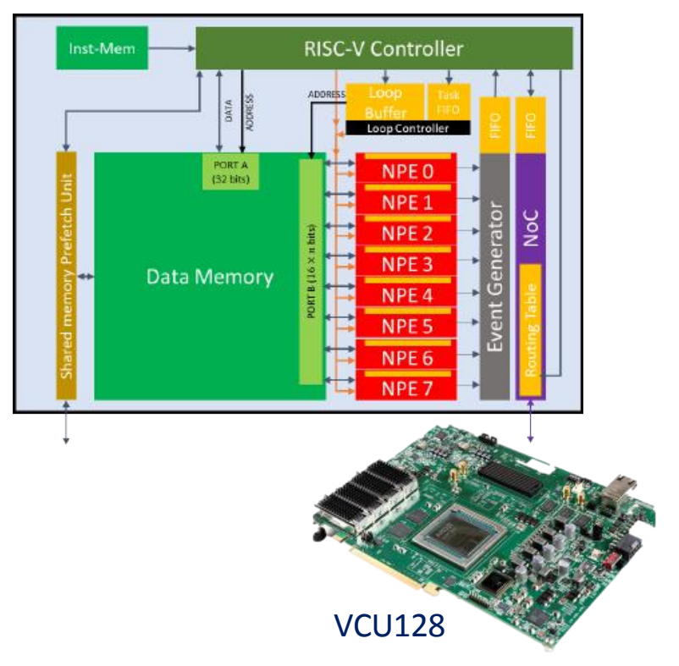


SENECA Neuromorphic Accelerator

FPGA-3 Neuromorphic Accelerator

- **Target**: Multi-core neuromorphic architecture on the FPGA for event-based neural network. To be evaluated in the context of object detection.
- **FPGA**: Virtex UltraScale+ HBM VCU128
- **Memory**: 2Mb per core
- **Cores**: 32
- **NN type**: Spiking Neural Network, Recurrent Neural Network, Hybrid Neural Network, Convolutional Neural Network, Object Detection Neural Network (YOLO).
- **Features**: Event-based processing, flexible neuron and learning support, depth-first convolution.
- **Status**: FPGA implementation validated.
- **Use cases**: UC4.1 and UC4.2 object detection in X-ray and ultrasound images/video

FPGA and single core architecture of SENECA

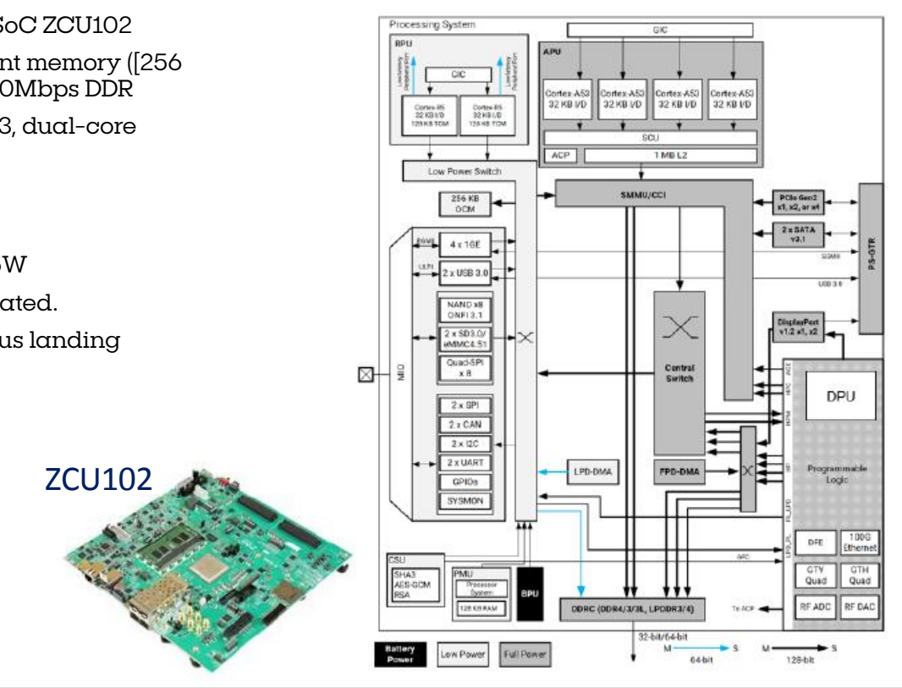


Accelerated inference of DNN for runway detection

FPGA-4: Runway detection

- **Target**: runway detection during landing manoeuvres
- **FPGA**: Xilinx Zynq UltraScale+ MPSoC ZCU102
- **Memory**: PL 512MB DDR4 component memory ([256 Mb x 16] devices) at 1200MHz / 2400Mbps DDR
- **Cores**: quad-core Arm® Cortex-A53, dual-core Cortex-R5
- **NN type**: Modified Yolo V4
- **Accuracy**: ~ 90%
- **Expected power consumption**: ~35W
- **Status**: FPGA implementation validated.
- **Use cases**: UC3.4 Robust autonomous landing

FPGA Architecture of Run Way Detection



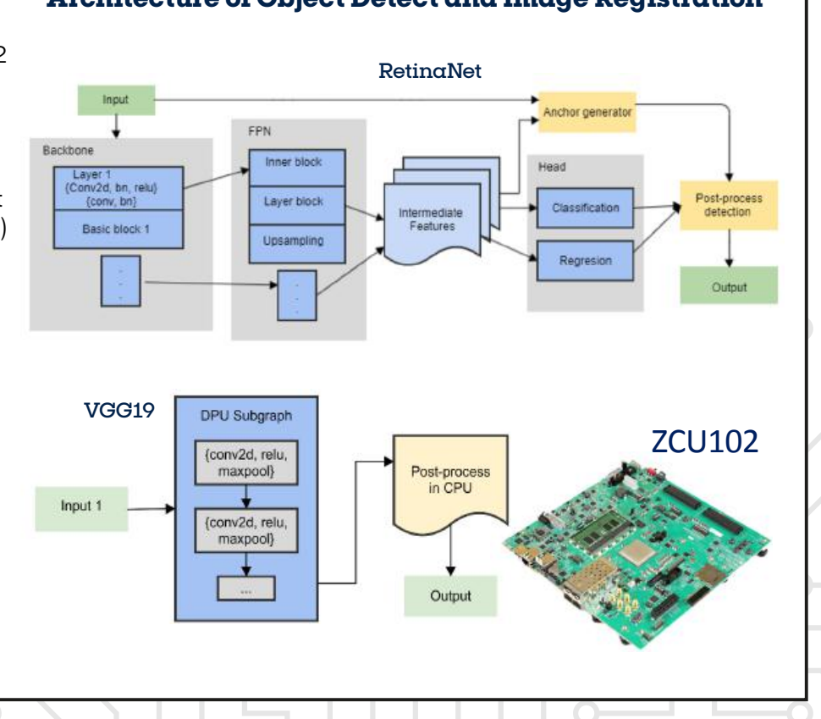
Accelerated inference of DNN for object detection and image registration

FPGA-5 Object detection and image registration

- **Target**: Object detection, and image feature extraction

FPGA: Xilinx Zynq UltraScale+ MPSoC ZCU102
Memory: Mem Bandwidth: 1.073 MB/s , Mem IO: 298 MB
Cores: RetinaNet: 3, VGG19: 2
NN type: RetinaNet and VGG19
Accuracy: RetinaNet mAP50: 0.53, for VGG19 it does not apply (is not a classification network)
Expected power consumption: RetinanNet: 39.97 W, VGG19: 26.33 W
Status : FPGA implementation validated.
Use cases: UC3.4 Robust autonomous landing

Architecture of Object Detect and Image Registration

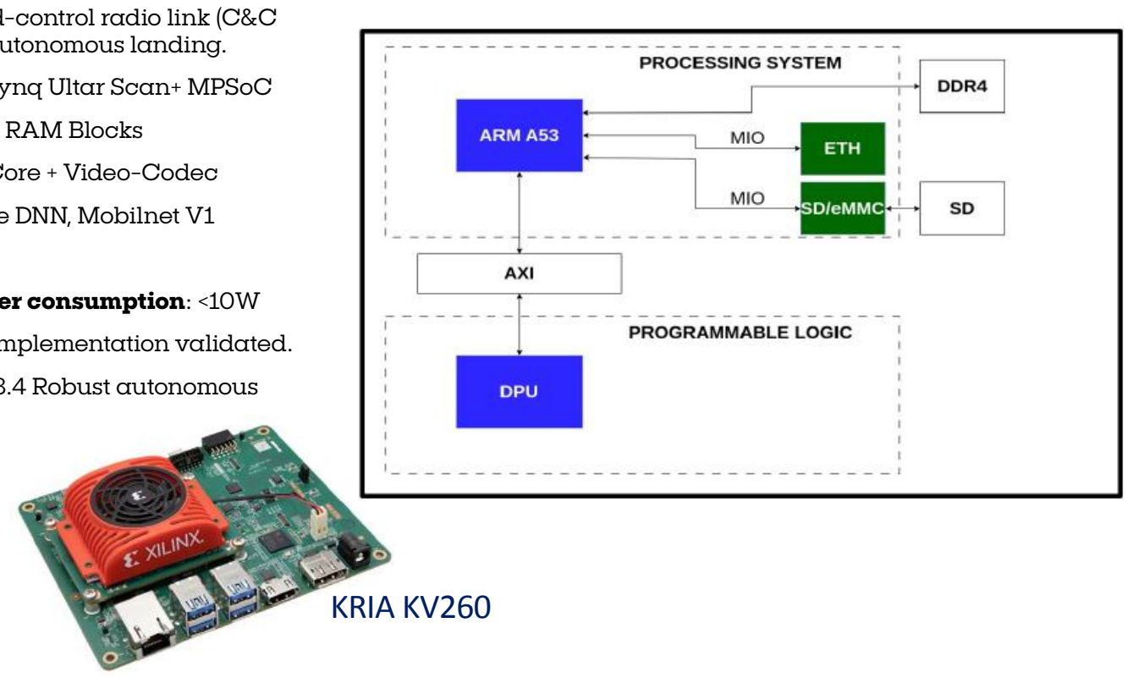


Accelerated inference of DNN for process communication parameters

FPGA-6 Command & control radio link

- **Target**: Prediction of the status of the command-and-control radio link (C&C RL) in Robust autonomous landing.
- **FPGA**: Xilinx Zynq Ultar Scan+ MPSoC
- **Memory**: 208K RAM Blocks
- **Cores**: Quad-Core + Video-Codec
- **NN type**: Dense DNN, Mobilnet V1
- **Accuracy**:
- **Expected power consumption**: <10W
- **Status**: FPGA implementation validated.
- **Use cases**: UC3.4 Robust autonomous landing

Architecture of Command & Control Radio Link



Conclusion

Twelve of the fourteen designs are available, two others are in the final development phase. The first results show very good energy efficiency. A complete validation, in the associated use case, has already been carried out for some of them and for the others the test and evaluation are ongoing.
These hardware solutions will be the basis of future highly energy efficient components and devices for Edge applications.

ANDANTE

AI for New Devices And Technologies at the Edge

andante-ai.eu   linkedin.com/company/andante-ai   mario.diaznava@st.com (coordinator)

“This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement № 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Germany, Netherlands, Portugal, Spain and Switzerland”

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

Scan Me
to visit website