ANDANTE SOC2.1 - VISAGE 2

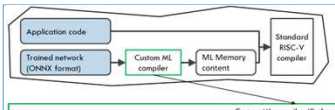
Overview

- CSEM's ***Visage*** ML SoC series target ULP system-on-chip (SoC) solutions that enable hierarchical processing of machine-learning inference, scalable from sub-mW power consumption to more than 1 TOPS/W efficiency at high throughput for different scenarios.
- ***Visage*** aims for moving towards computer vision at the extreme edge, where the computational complexity challenges the strict energy constraints of miniaturized and mobile devices.

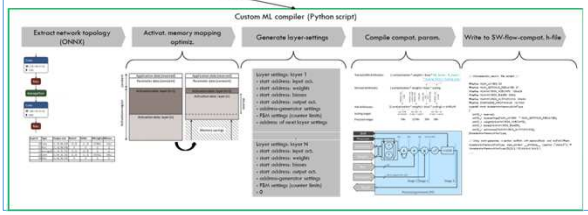
Feature	Visage1	Visage2
Total Memory Size	1.2MB	4MB
NVM Storage	Off-chip	MRAM + Off-chip
Always-on Detection Engine	Y	Y
Conv. Neural Network Acc.	Y	Y
# Clusters	1	4
Sparsity Exploitation	-	Y
Selective Execution Support	-	Y
Domain power gating	-	Y
Memory power gating	512 kB granularity	Bank granularity

Use-case 5.2
Vision-Based Human Computer Interaction

Tools & Methodology



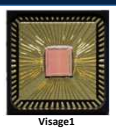
ML compiler



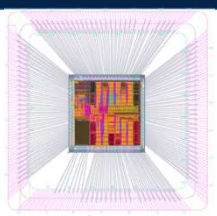
Flows (Initiated within ANDANTE):

- Quantization-aware training
- Mapping: Python-based ML compiler
  - Converting ONNX-format input file into a binary file that is loaded into the memory along with the application code.
- Verification and validation

Technology

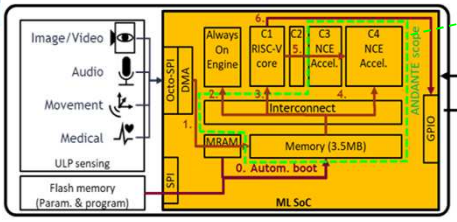


Visage1
(Previous Generation)



Visage2

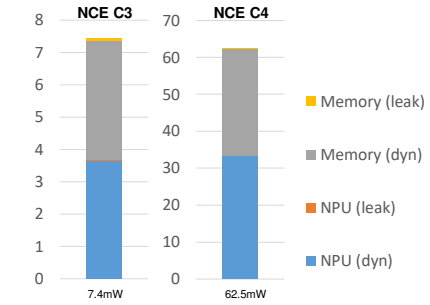
Tech: GF 22nm FDX



Visage2 ML inference SoC with 2 NCE blocks for ML acceleration

	C3 NCE [a.k.a. NPL]	C4 NCE [a.k.a. NPH]
Number of mem. Ports	2	4
Number of proc. Elements	4 (5x5 array)	16 (5x5 array)
MAC Units per cluster	100	400
Weight reuse	Local buffer	
Activation reuse	Systolic array + input sharing	
Precision support	8b, 16b [A+W]	
Base layer support	Convolution (Conv.), DS-Conv (Depth-wise separable Conv.), Dense, Min./Max./Avg.-Pooling, Scaling	
Maximum clock frequency	100 MHz	250 MHz

Results



Power consumption mW @ peak throughput

Model and dataset from UC5.2:

- LeNet-based 13-layer CNN
- 6-class Glance Detection

MAC precision	TOPS/W (block)	TOPS/W (system)
16b	5.9	3.2
8b	10.9	6.1

Peak Throughput	C3	C4
GOPS	20 (16b)	200 (16b)

A sub-mW dual-engine ML inference system-on-chip for complete end-to-end face-analysis at the edge.“, 2021 10.23919/VLSICircuits52068.2021.9492401

A Construction Kit for Efficient Low Power Neural Network Accelerator Designs, 2022, https://doi.org/10.1145/3520127

An Ultra-Low-Power Serial Implementation for Sigmoid and Tanh Using CORDIC Algorithm, 2023, 10.23919/DATE56975.2023.10136960

Impact

- Follow-up industrial and EU projects
- Demonstrators for fairs, events, and customer meetings
- CSEM's IP Library for Edge ML

Progress beyond SoA

- End-to-end ML inference at the edge with hierarchical computing
- At-par performance, while providing higher flexibility / flexible performance-energy scaling

Lessons learned

- Heterogeneous computing platforms with dedicated accelerators (e.g., Visage) provide scalability and flexibility, which are key to keeping up with fast-moving application trends.

ANDANTE

AI for New Devices And Technologies at the Edge

andante-ai.edu 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”

Scan Me to visit website