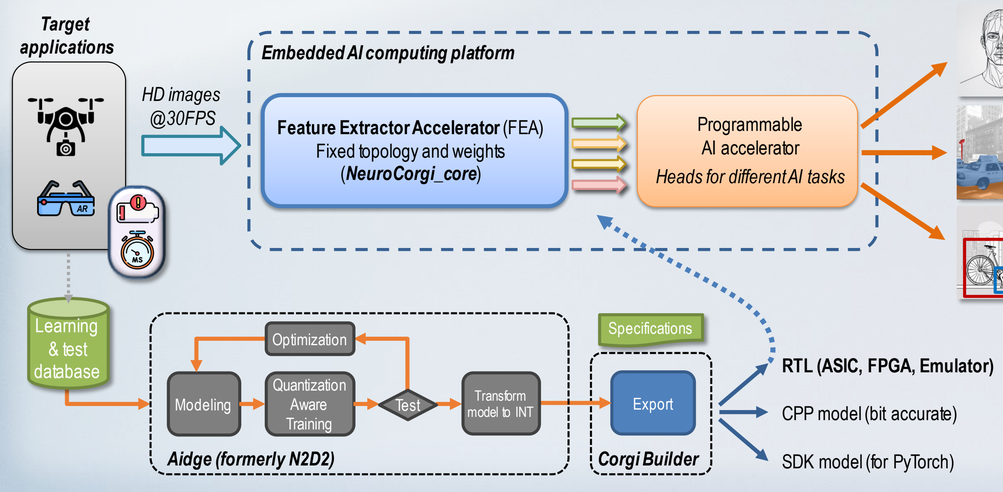
A 772µJ/frame ImageNet Feature Extractor Accelerator on HD Images at 30FPS

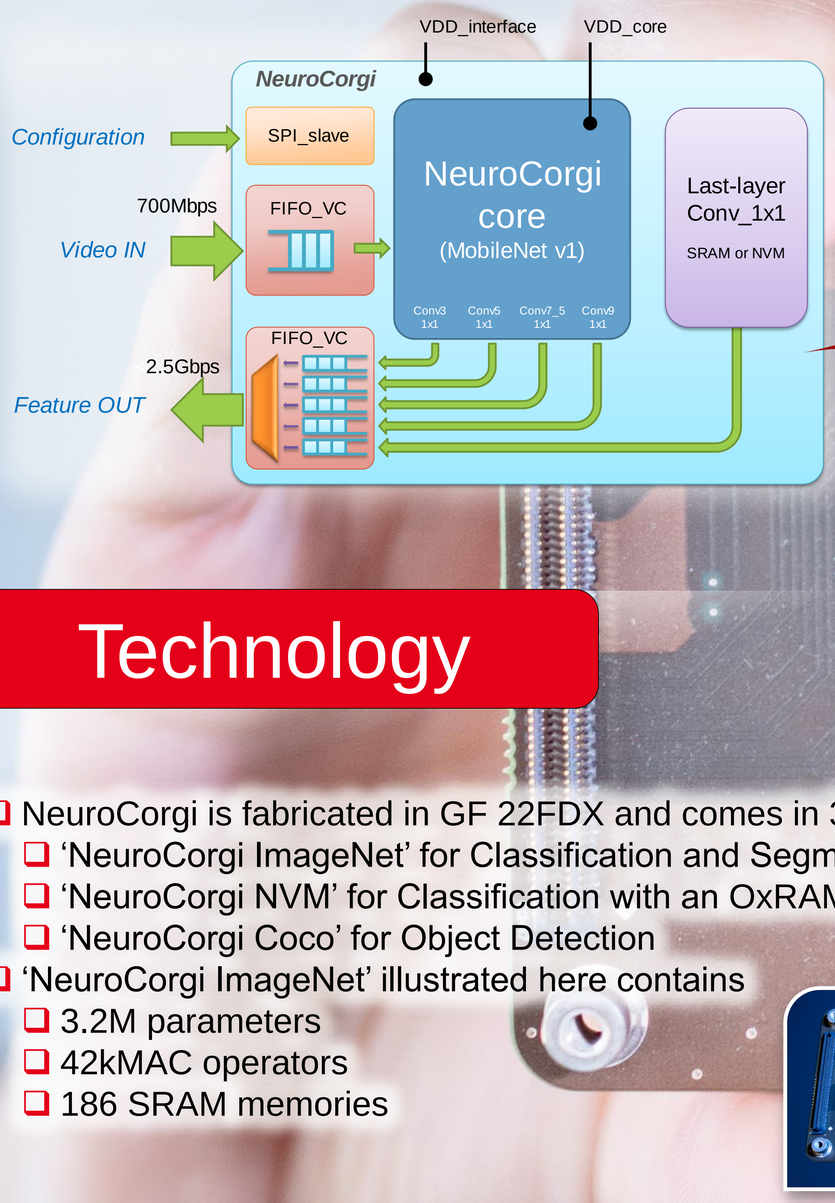
Ivan Miro-Panades[1], Vincent Lorrain[2], Lilian Billod[1], Inna Kucher[2], Vincent Templier[2], Sylvain Choisnet[1], Nermine Ali[2], Baptiste Rossigneux[2], Olivier Bichler[2], Alexandre Valentian[1]
[1]CEA, List, Grenoble, France
[2]CEA, List, Palaiseau, France

Overview

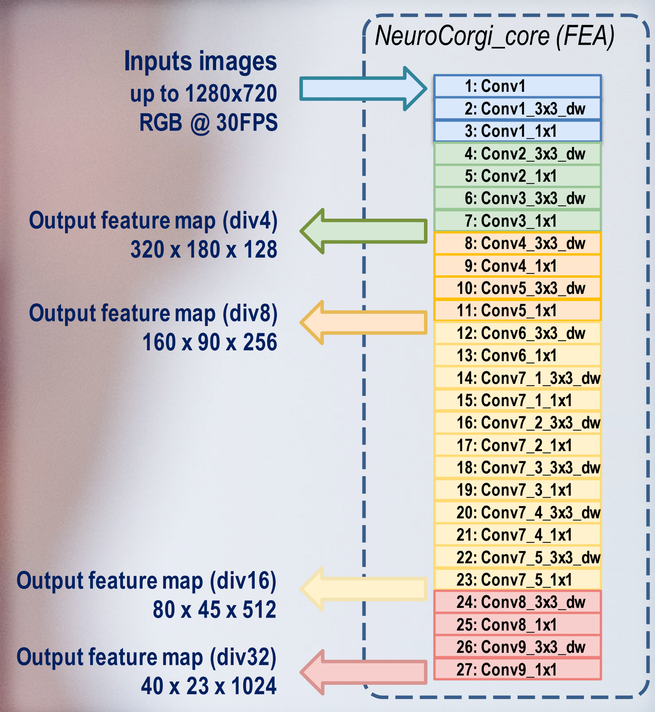


- NeuroCorgi is a Feature Extractor Accelerator (FEA) circuit
- Its Key design principle draws inspiration from the Human Brain, which shows that generic Feature Extractors can be used across a range of applications
- Leveraging a fixed MobileNet-V1 backbone and the Transfer Learning concept, its power dissipation beats the best State-of-the-Art accelerators by 9× to 1000×
- Built in 22FDX, it addresses Classification, Semantic Segmentation and Multi-Object Detection applications
- Designed using a set of tools for training the model (Aidge) and generating a complete new circuit (CorgiBuilder) for a given AI model

Architecture

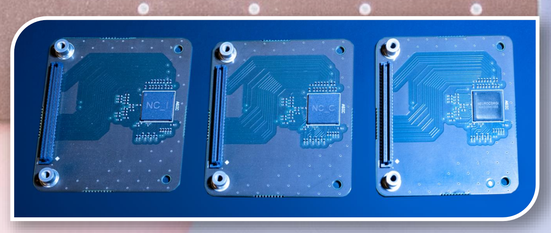


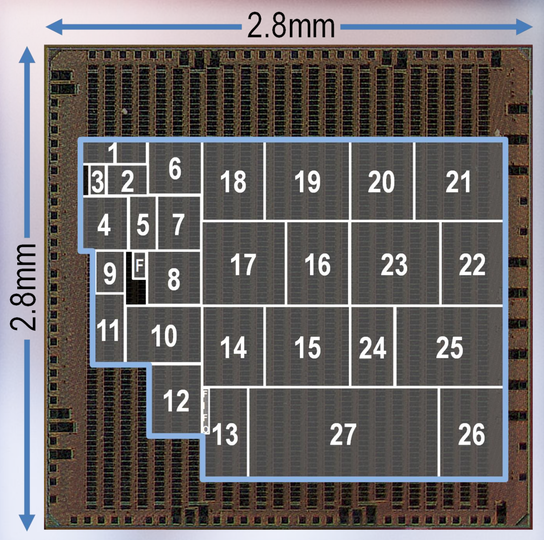
- Streaming architecture
- Support any image size up to HD (1280x720)
- MobileNet-V1 topology (27 layers)
- Trained on 4b weights and activations
- Accuracy of 70.42% on ImageNet
- 73.3 mIoU on Cityscapes semantic segmentation



Technology

- NeuroCorgi is fabricated in GF 22FDX and comes in 3 variants
  - 'NeuroCorgi ImageNet' for Classification and Segmentation
  - 'NeuroCorgi NVM' for Classification with an OxRAM-based classifier
  - 'NeuroCorgi Coco' for Object Detection
- 'NeuroCorgi ImageNet' illustrated here contains
  - 3.2M parameters
  - 42kMAC operators
  - 186 SRAM memories

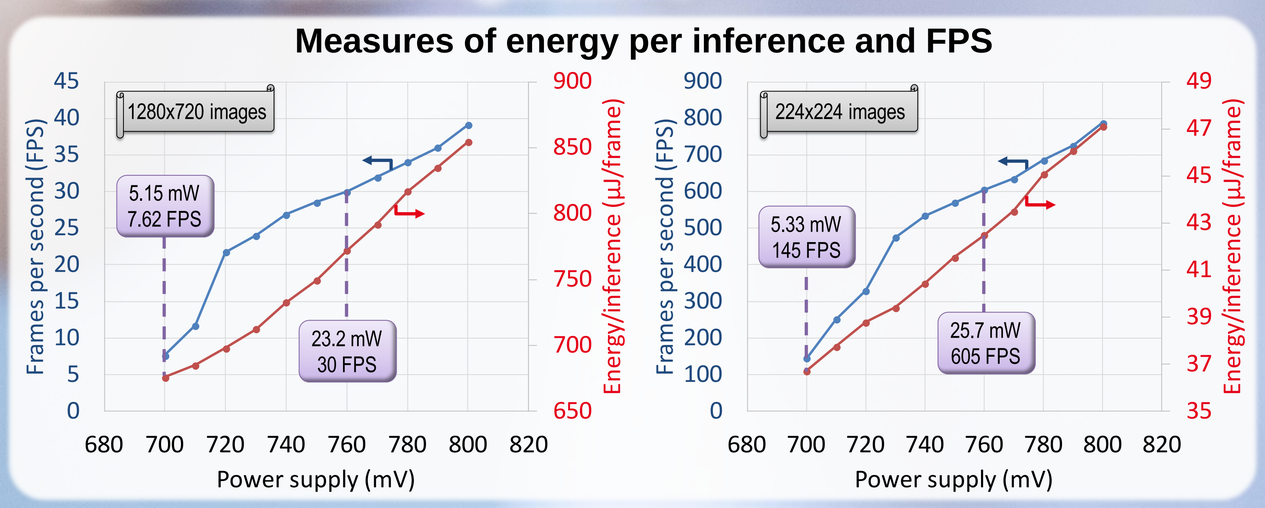




Chip summary	
Technology	GF 22FDX
Chip area	7.86mm²
FEA area	4.45mm²
# multipliers	42k
# SRAM memories	186
SRAM memory	1.1MB
Main clock	59MHz
AI model	MobileNet v1
Training dataset	ImageNet
Batch size	1

Results

- 'NeuroCorgi ImageNet' measurements show outstanding results
- End-to-end energy efficiency: 30.9 TOPS/W (4b weights and activations)
- At least 9× better energy per inference than the State-of-the-Art
- Ultra-low power consumption:
  - 23.2mW for 1280x720 images at 30FPS
  - 1.5mW for 224x224 images at 30FPS



			ISSCC'20 [1]	JSSC'23 [2]	JSSC'23 [3] DIANA	JSSC'24 [4] Marsellus	JSSC'24 [5] DynaPlasia	ISSCC'22 [6] Hiddenite	This work [7] NeuroCorgi
Technology			12nm	4nm	22nm	22nm FDX	28nm	40nm	22nm FDX
Application			Server	Mobile	Edge	AI-IoT	Embedded	Embedded	Embedded
Area (mm²)			709 (chip)	4.74 (core)	3.3 (core)	18.7 (chip) 2.42 (core)	20.25 (chip)	9 (chip) 4.36 (core)	7.86 (chip) 4.45 (FEA)
Programmability			Programmable	Programmable	Programmable	Programmable	Programmable	Configurable	Fixed
Power consumption (mW)			25W – 276W	381 – 5133	–	12.8 – 123	261	85.4 – 534.7	5 – 37
ImageNet use case	**Training dataset**		ImageNet	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet
	AI model		ResNet50 v1	MobileNet TPU	ResNet18	ResNet18	ResNet18	ResNet50	MobileNet v1
	Precision (bits)		8	8	Analog + digital	RBE 4x4b	9w, 8a	ternary(w), 8a	4w, 4a
	Top-1 accuracy (%)		74.93	–	64.1	68.5	70.4	70.09	70.42
	Inferences/second (FPS)	224x224	78563	3433	277	20.8	776[γ]	169.7[α]	788
		1280x720	–	–	–	–	–	–	39
	FEA latency (ms)	224x224	0.2[βδ]	0.29[δ]	3.6[δ]	48[δ]	1.29[γδ]	5.92[αδ]	1.23
		1280x720	–	–	–	–	–	–	6.90
	TOPS/W	224x224	4.14	11.59	5.52[α]	5.83	10.8[γ]	16[α]	30.9[ϕ]
		1280x720	–	–	–	–	–	–	30.9[ϕ]
	Best FEA Energy/inference (µJ/frame)	224x224	2000[δ]	340[δ]	659[αδ]	557[δ]	336[γδ]	503[αδ]	36.7
		1280x720	–	–	–	–	–	–	676

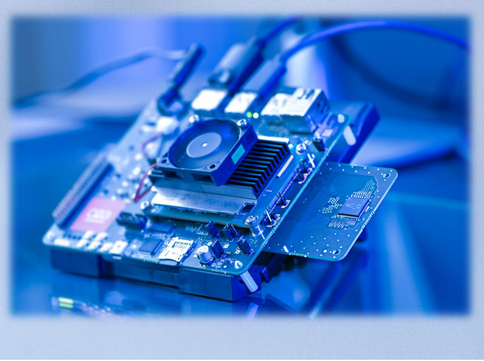
1 MAC = 2 Ops. Zero skipping included as MACs [α]Without considering off-chip memory accesses. [β]Latency reported on Inception v3. [ϕ]Only feature extraction (no FC layer)
[δ]Estimated feature extraction part. Assuming 1Conv OP = 1FC OP for latency and energy. FC is 0.18% (0.03%) of total MACs/frame on 224x224 images for MobileNet v1 (ResNet18)
[γ]Power & latency off-chip weight loading from external DRAM / external host CPU / on-chip network / on-chip memory access / refresh are not included.

Acknowledgments

ANDANTE
(ECSEL 876925)
EU project

PREVAIL
(DIGITAL 101083307)
EU project

Reference:
[1] Y. Jiao et al., International Solid-State Circuits Conference, 2020.
[2] J. -S. Park et al., Journal of Solid-State Circuits, 2023.
[3] P. Houshmand et al., Journal of Solid-State Circuits, 2023.
[4] F. Conti et al, Journal of Solid-State Circuits, 2024.
[5] S. Kim et al., Journal of Solid-State Circuits, 2024.
[6] K. Hirose et al., International Solid-State Circuits Conference, 2022.
[7] I. Miro-Panades et al., Asia Pacific Conference on Circuit and Systems, 2024.



Scan Me
to visit website