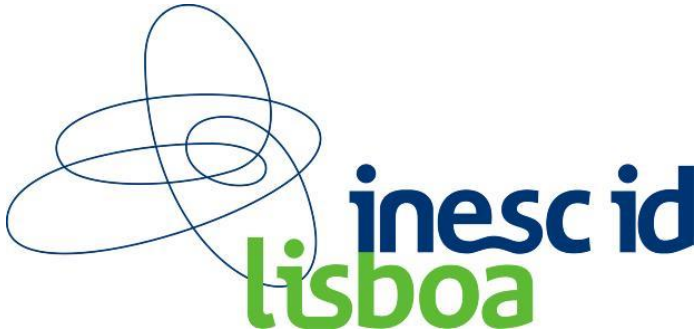
technology
from seed

A Full Featured Configurable Accelerator for Object Detection With YOLO

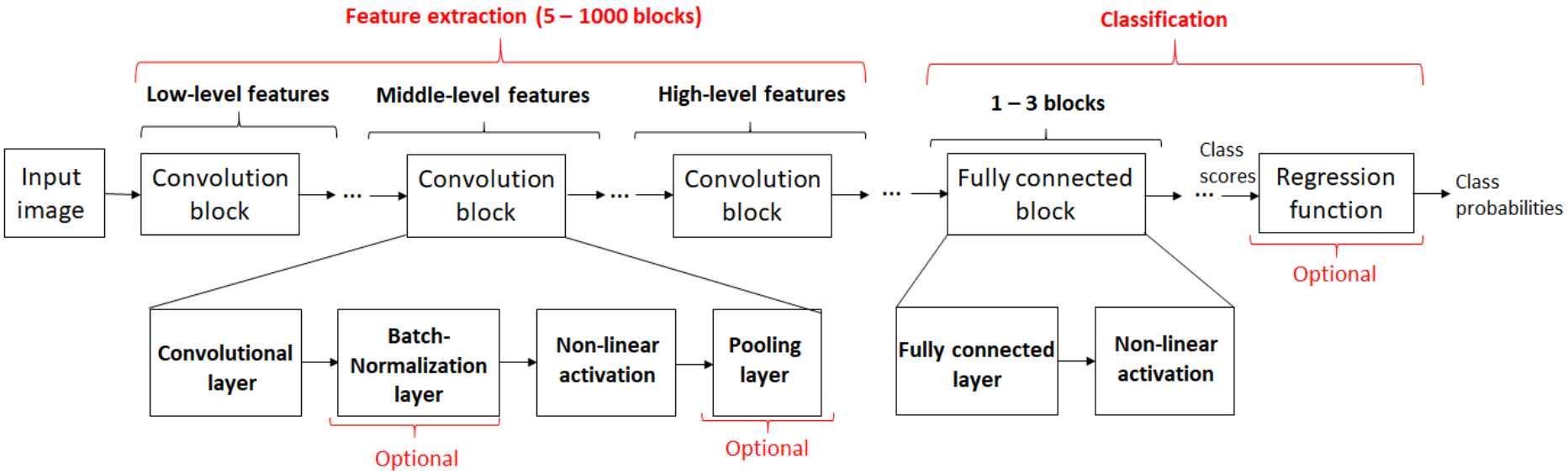
Daniel Pestana, Pedro Miranda, João Lopes, Rui Duarte
Mário Véstias, Horácio Neto, José T. de Sousa



Objectives

- Design a configurable IP core to execute na object detector system based on YOLO
- Design complete algorithm, including pre- and post-processing
- Based on configurable IP cores that support any version of YOLO
- Verilog IP core portable to FPGA devices from all vendors and Application-Specific Integrated Circuits (ASICs)

Convolutional Neural Networks

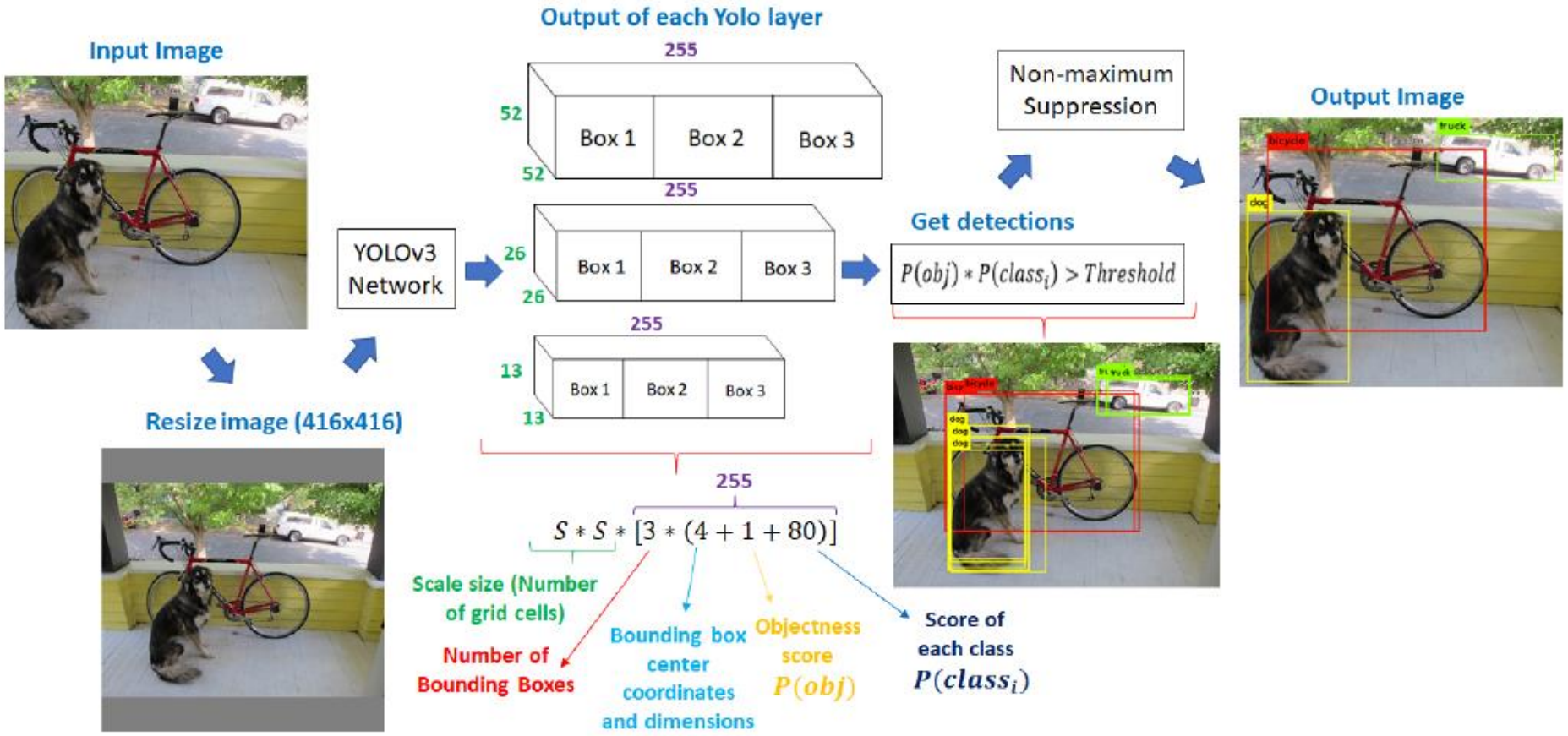


Complexity of Convolutional Neural Networks

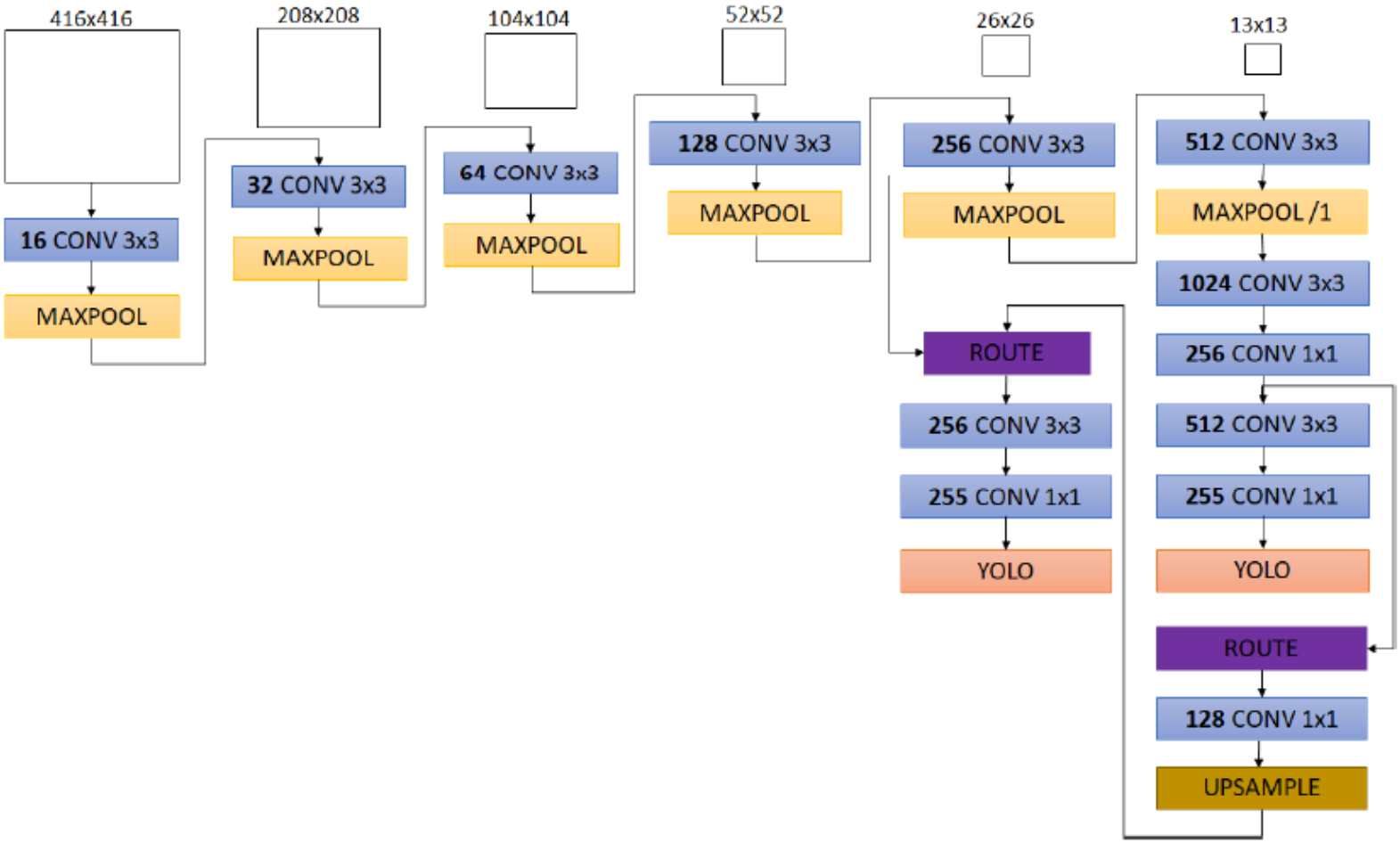
Model	# Million Parameters	# GMAC
AlexNet	60	0.65
VGG16	138	7.80
ResNet-101	40	3.80
ResNet-152	55	5.65
Darknet-53	62.2	32.90
CSPDarknet-53	27.6	26
Darknet-53-Tiny	8.8	2.78
CSPDarknet-53-Tiny	6.5	3.75

Tiny versions reduce considerably the memory and computational requirements

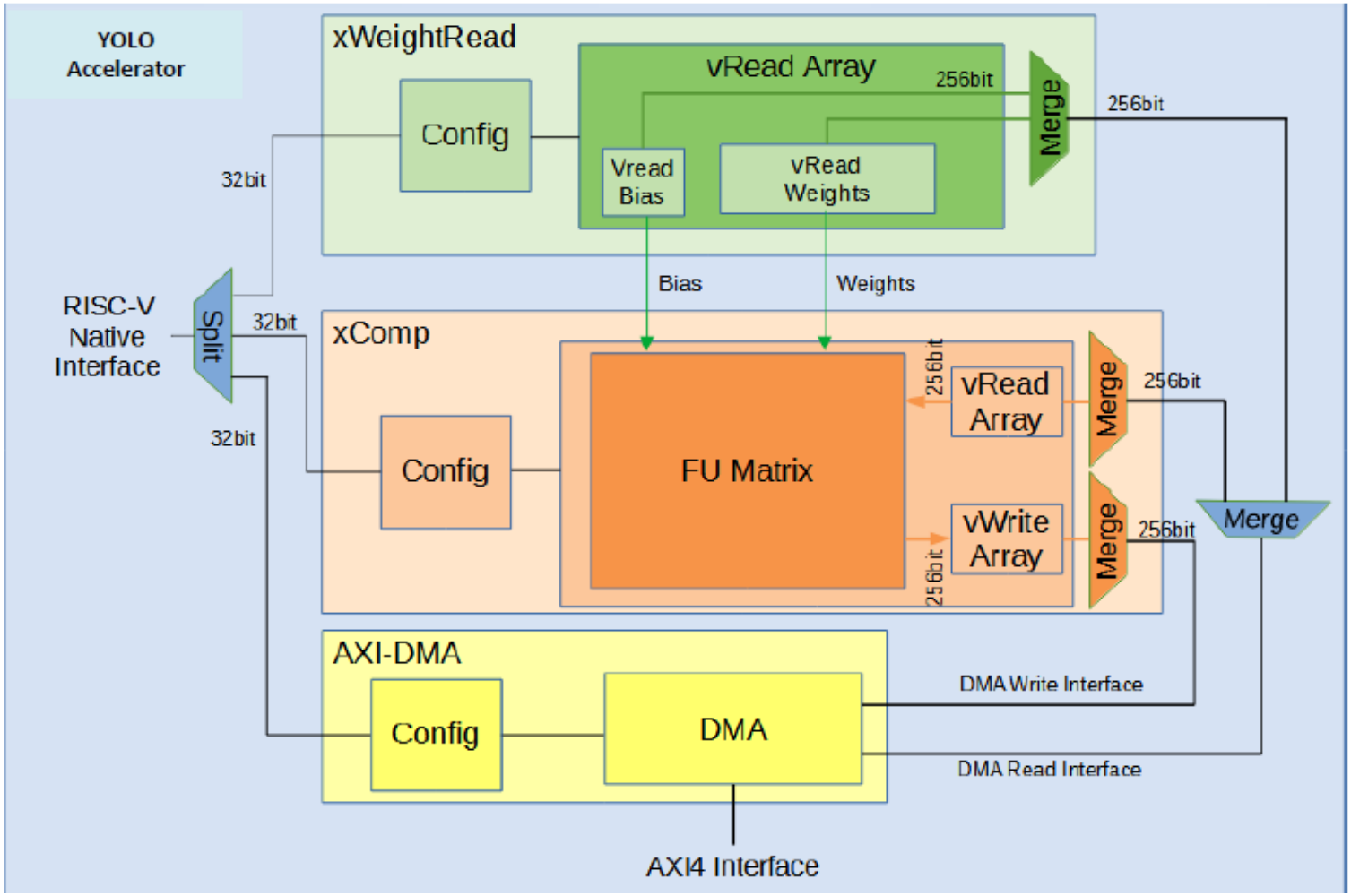
Object Detection with YOLO



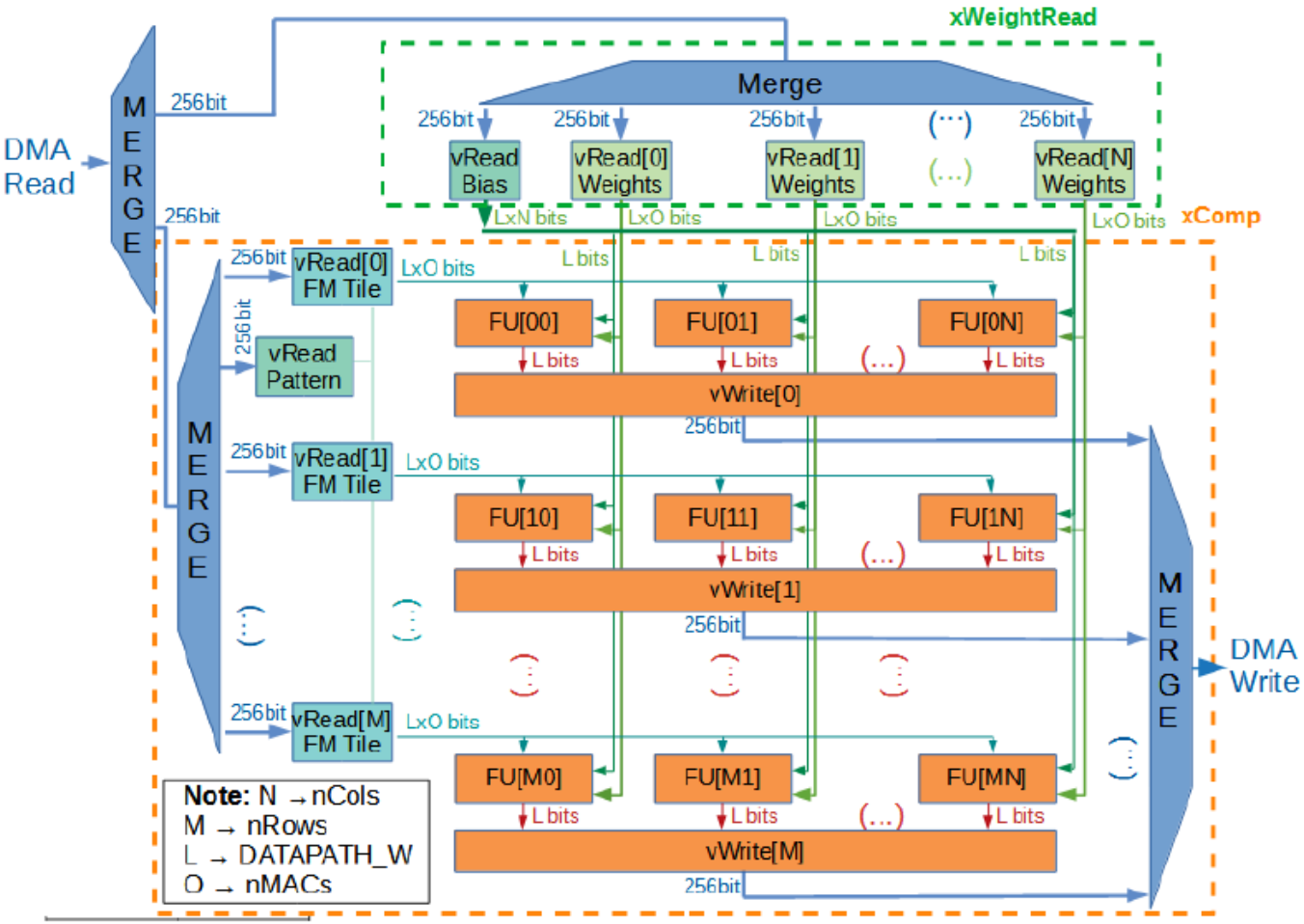
Architecture of YoloV3-Tiny



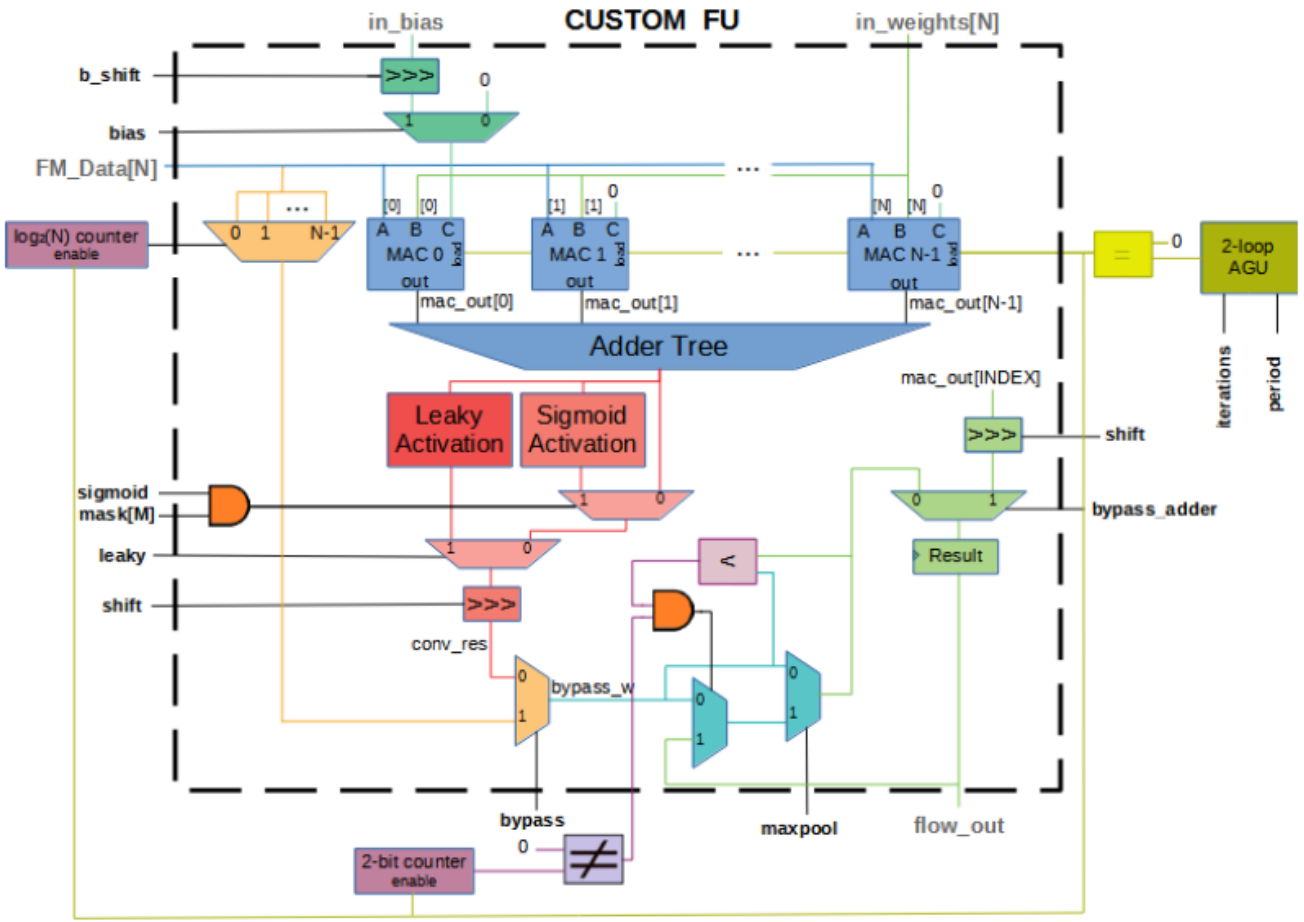
High-Level Architecture of the Accelerator



Architecture of the Main Unit – FU Matrix



Architecture of the Core unit



Two versions of YOLO were implemented

Parameter	Tiny-YOLOv3	Tiny-YOLOv4
nCols	16	24
nRows	13	13
nMACs	4	4
DDR_ADDR_W	32	32
DATAPATH_W	16	16
VREAD_TILE_EXT_ADDR_W	15	15
VREAD_BIAS_ADDR_W	3	3
VREAD_WEIGHT_ADDR_W	14	14
VREAD_TILE_ADDR_W	15	15
VREAD_PATTERN_ADDR_W	10	10
VWRITE_ADDR_W	8	8

Results - Area

The two implementations were implemented and tested on FPGA

Component	BRAM	FF	LUT	DSP
AXI Interconnect	0	9,887	3,442	0
DDR4 Controller	25.5	11,918	9,697	3
RISC-V CPU	0	902	2,569	4
Internal memory	17	41	60	0
AXI Cache	1	592	629	0
YOLO IP core v3	339	86,319	103,655	832
YOLO IP core v4	403	124,761	146,820	1248
Ethernet	1	382	193	0
UART	0	89	86	0
Timer	0	130	2	0
Others	0	728	480	0
Total v3	383.5	110,988	138,946	839
Total v4	447.5	149430	182,111	1255

Results - Performance

YOLOv3-Tiny

Platform	Time (ms)	FPS
CPU (Intel i7-8700 @ 3.2 GHz)	828.3	1.2
GPU (RTX 2080 Ti)	15.4	64.9
FPGA (SoC-YOLO)	30.9	32.4

YOLOv4-Tiny

Platform	Time (ms)	FPS
CPU (Intel i7-8700 @ 3.2 GHz)	1054.1	0.9
GPU (RTX 2080 Ti)	19.7	50.7
FPGA (SoC-YOLO)	32.1	31.2

Comparison with SoA

	[38]	[39]	[40]	SoC-YOLO
FPGA	UltraScale+ XCZU9EG	Virtex-7 XC7VX485T	Zynq 7020	UltraScale XCKU040
Freq.(MHz)	-	200	100	143
LUT (K)	-	49	26	139
BRAM	-	70	93	384
DSP	-	2304	160	839
FPS	104.2	-	1.9	32.4
FP (bits)	8	18	16	16
GOPS	-	-	10.5	180
MOPS/s/kLUT	-	-	403.8	1295.0
MOPS/s/DSP	-	-	66.3	215.5
Power (W)	-	4.81	3.36	3.87

Conclusions and Future Work

- A new configurable hardware accelerator for the execution of the Tiny versions of YOLO.
- The IP core consists of a matrix of vector functional units
- The IP core also accelerates the pre-CNN procedure and the drawing of the detections in the post-CNN procedure.
- The accelerator was integrated into a RISC-V-based SoC platform and then configured for real-time execution of YOLOv3-Tiny and YOLOv4-Tiny object detectors.

- As future work, the hardware accelerator can be improved with a more aggressive quantization.