JUNE 22-26, 2025 ◆ SAN FRANCISCO, CA

Flexible Integration of a Neural Network Library in TensorFlow Lite Framework for Efficient Programmable Near-Memory Computing Architectures

T. Bricout, M. Kooli, J.-P. Noel, M. Ramirez Corrales, H.-P. Charles
Univ. Grenoble Alpes, CEA, LIST F-38000 Grenoble

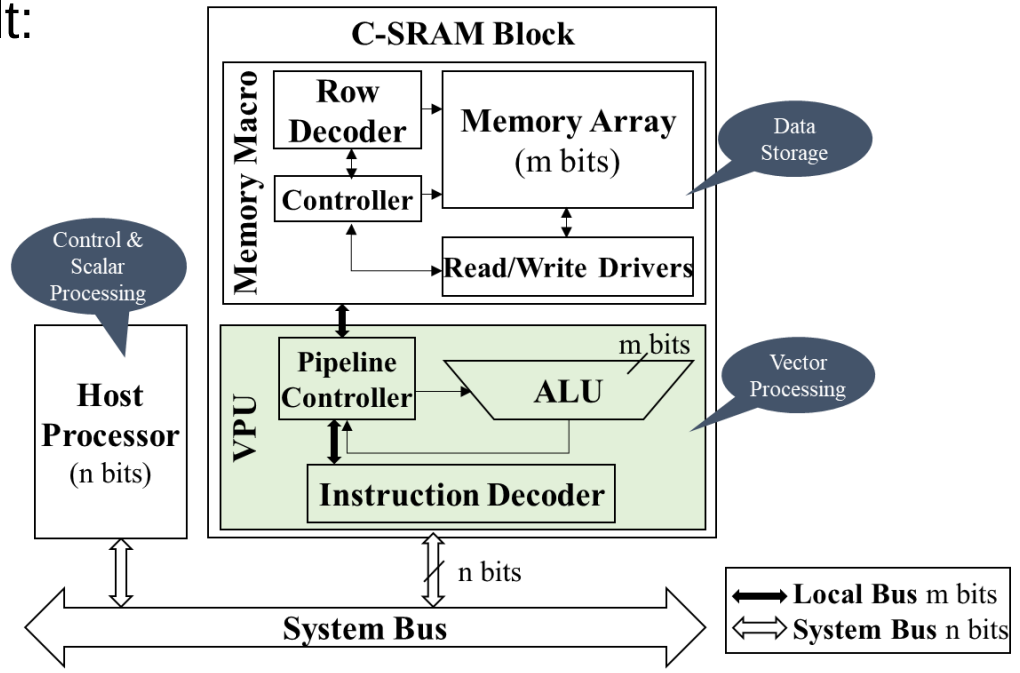
This work has been supported by NimbleAI project which has received funding from the EU's Horizon Europe Research and Innovation program (Grant Agreement 101070679), and by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Agreement 10039070).

Context & Motivation

Converting an application for an NMC target is difficult:

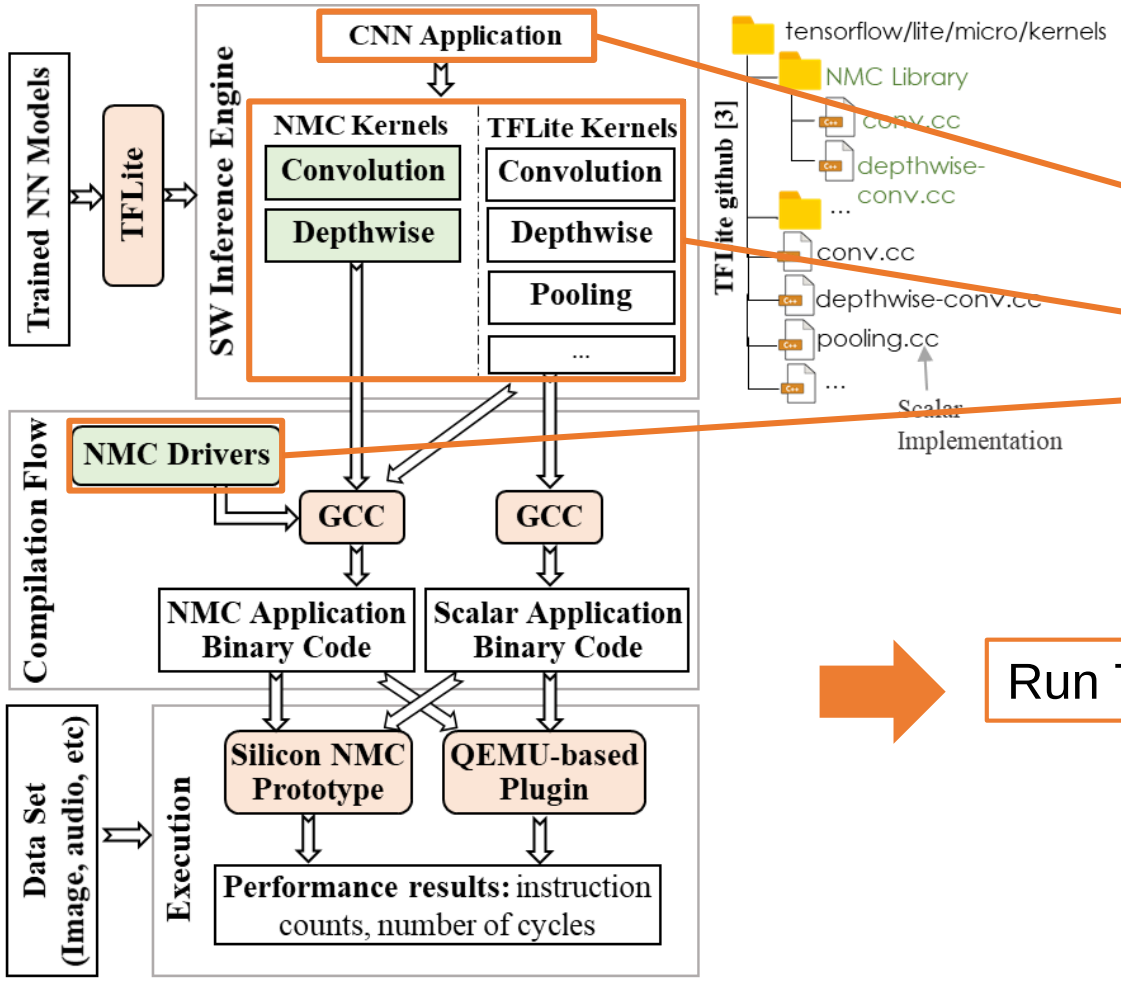
- Rethinking the way memory is used
- Addition of assembly-like instructions in C code
- Requires modification of each application

Category	Width (bits)	Mnemonic	Description
Memory	NMC line	copy	Copy a line into another
	8, 16, 32	bcast	Broadcast value in a memory line
	8, 16, 32	slli, srli	Shift left or right logical immediate
Logical	NMC line	and, or, xor	Logical AND, OR, XOR
	NMC line	nand, nor, xnor	Negation of logical AND, OR, XOR
Arithmetic	8, 16, 32	add, sub	Arithmetic addition and subtraction
	8, 16, 32	mul, mulhi	Arithmetic integer multiplication
	8, 16, 32	mac	integer multiply-accumulate



How can we support SW developers in running AI applications on optimized NMC architectures?

Main idea : Use TensorFlow Lite for Microcontroller to run AI model on NMC architecture

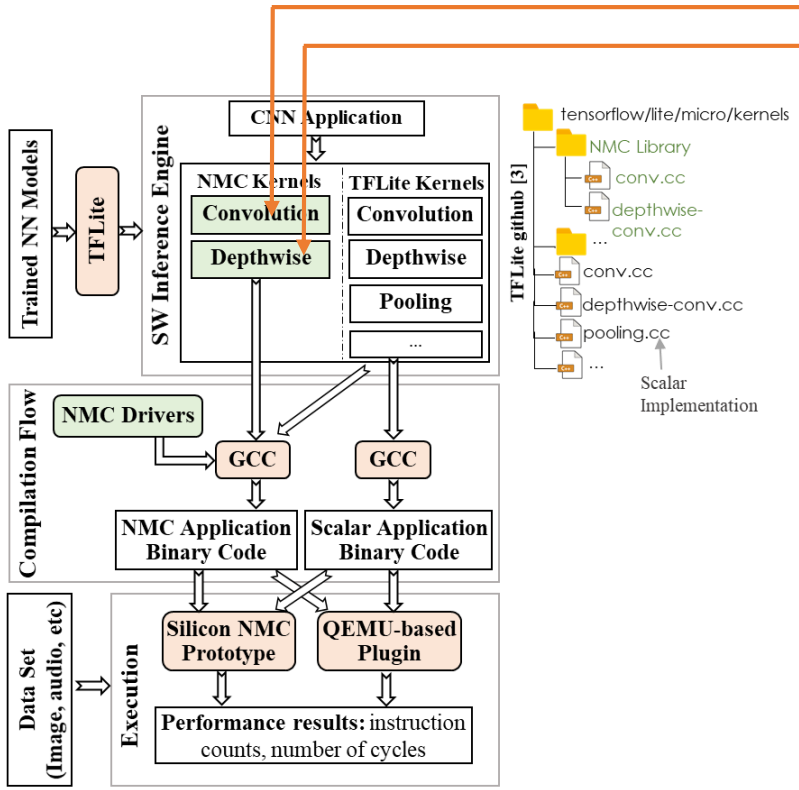


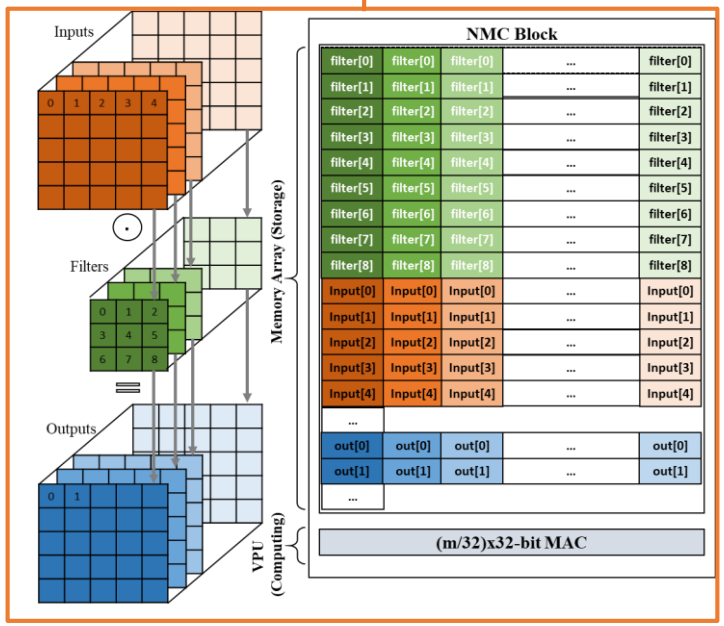
TFLM is a framework to run AI applications on microcontroller architectures with :

- Make application and AI model architecture agnostic
- Libraries of AI kernels for many architectures
- Specific integrated IP drivers
- Compilation script to adapt the application on the architecture with compilation flags

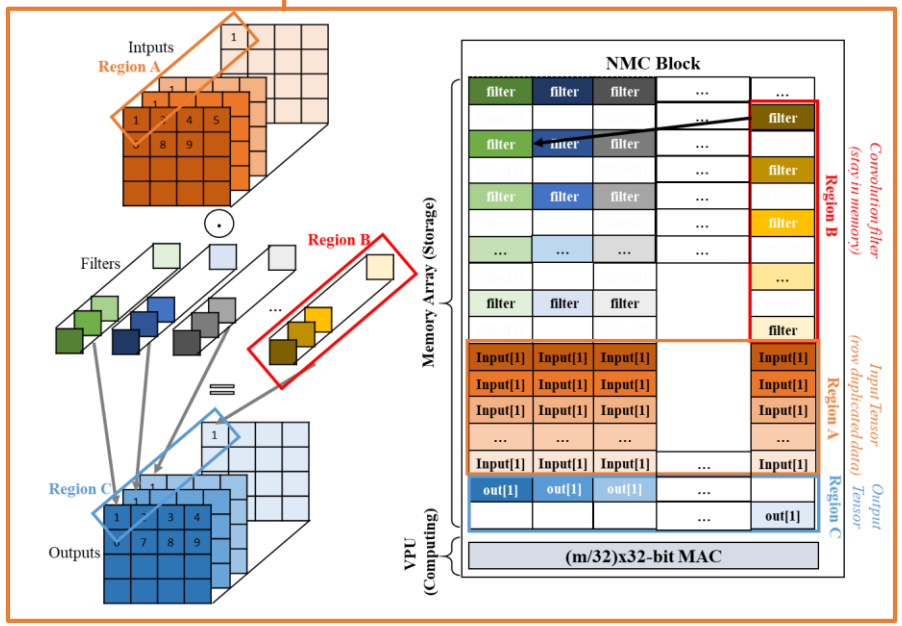
Run TinyML model on NMC architecture without modification

Integrate an NMC library in TensorFlow Lite for the Microcontroller Framework



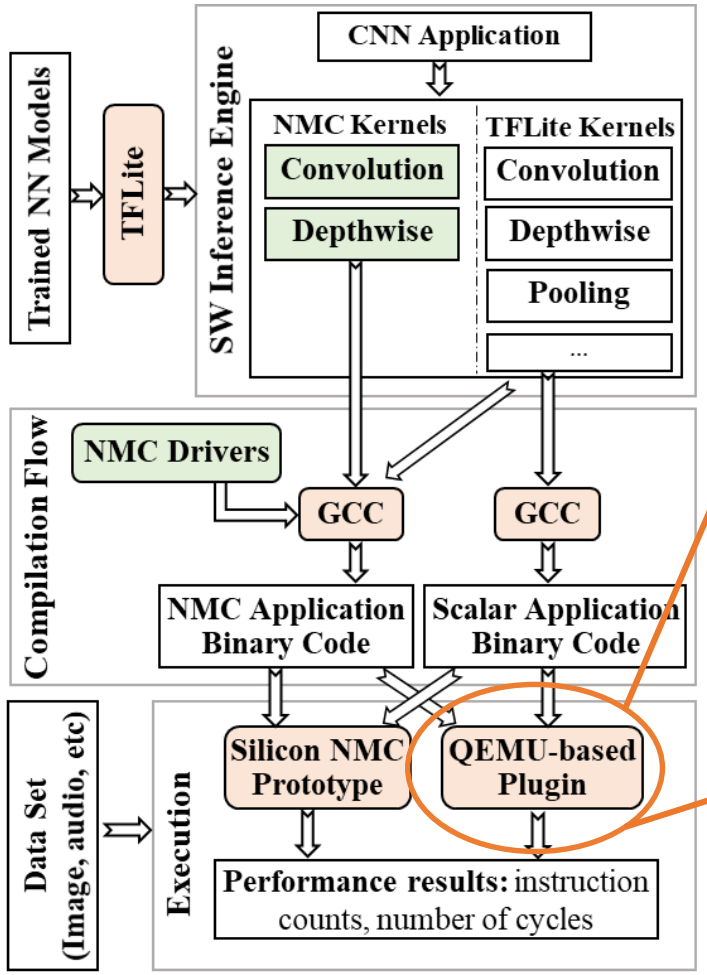


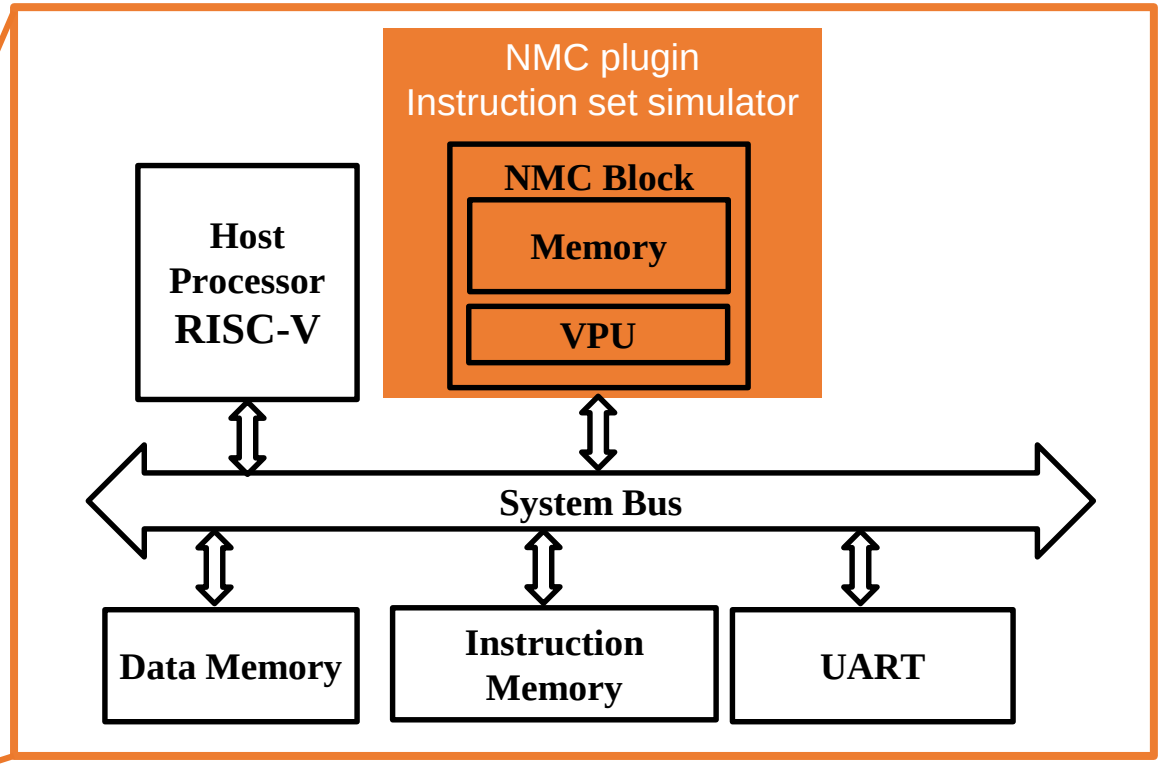
Depthwise Convolution Kernel



Convolution Kernel

Experimental setup

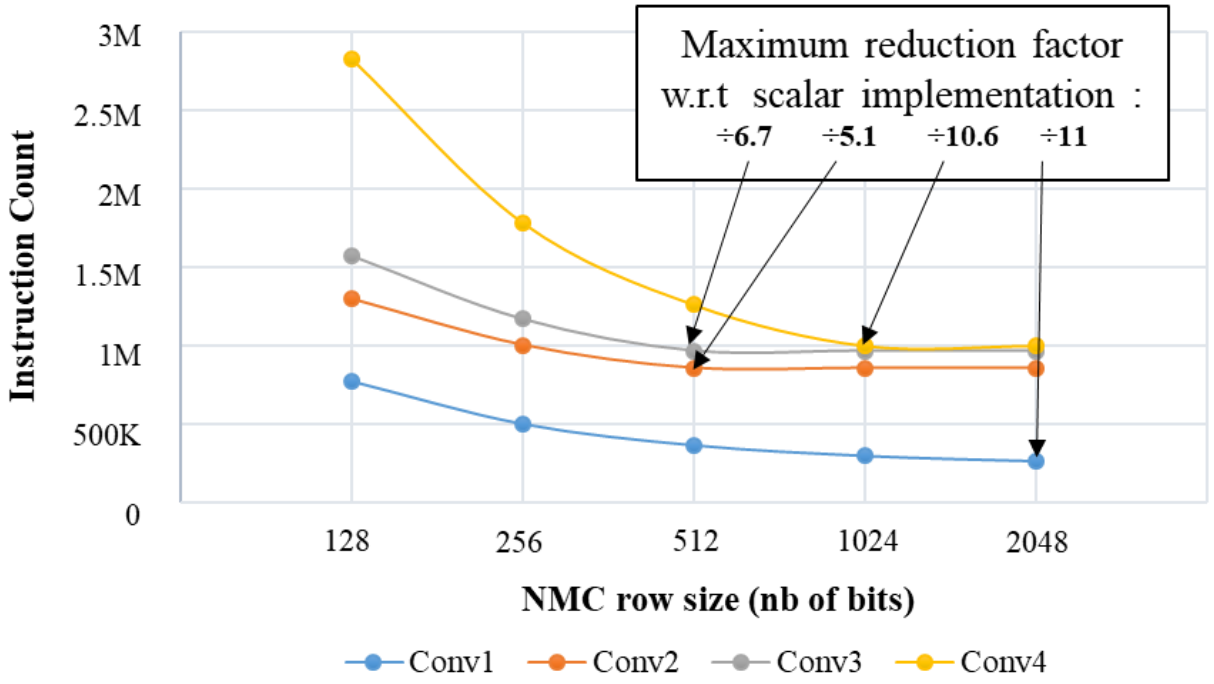




RISC-V architecture emulator

- Instruction level simulator
- Instruction counter integrated inside NMC plugin

Additional Content : Optimizing micro-architectural parameters with hardware/software co-design

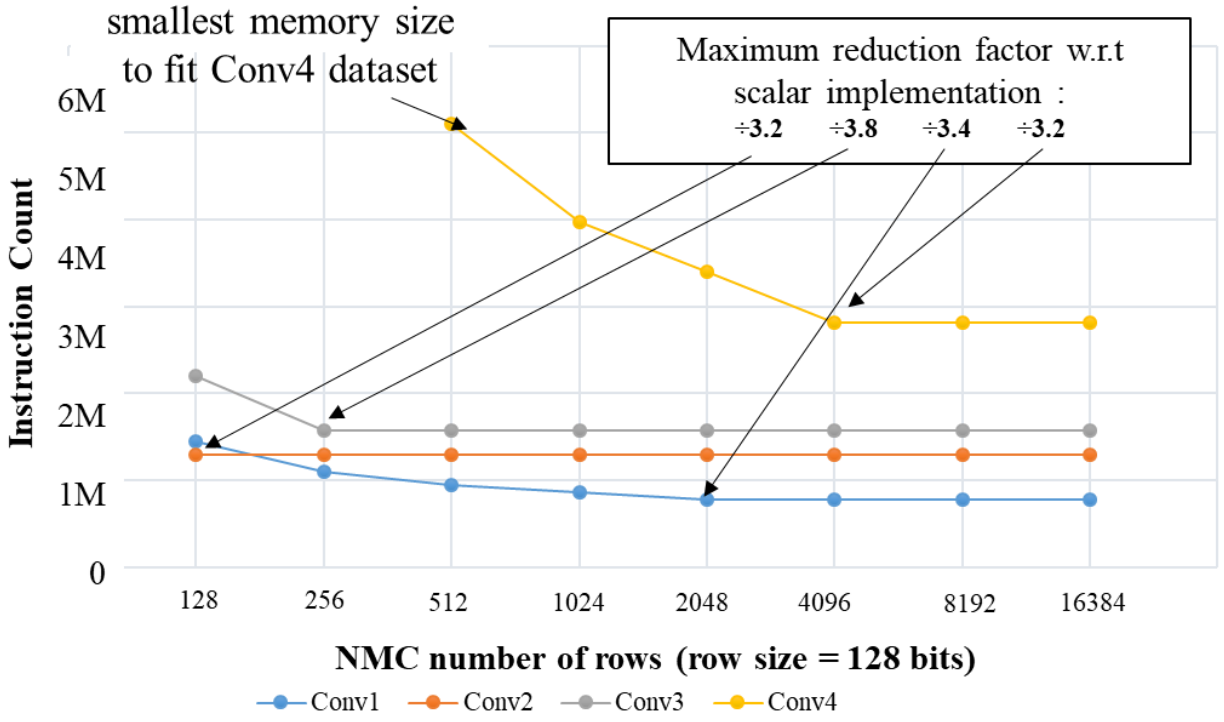


Variation of memory row size form 128 to 2048 bits:

- Memory size is fixed to 256 KBytes
- Change vector size for NMC operation from 4 to 64 32-bits word
- Convolution Kernel tested with 4 dataset

Determine the optimal memory size parameters using a QEMU-based architecture emulator

Additional Content : Optimizing micro-architectural parameters with hardware/software co-design



Variation in number of lines from 128 to 16384 :

- Memory row size is fixed to 128 bits (4 32-bits words)
- Memory size changed form 2 KBytes to 256 KBytes
- Convolution Kernel tested with 4 dataset

Determine the optimal memory size parameters using a QEMU-based architecture emulator

Evidence : Speed-up on a CNN inference application run on an optimal NMC architecture

Kernel	Data Set			Given Name	# Instructions		Reduction Factor
	Input	Filter	Output		NMC	Scalar	
Convolution	1*12*12*32	64*1*1*32	1*12*12*64	Conv1	771K	3.0M	x3.8
	1*48*48*8	16*1*1*8	1*48*48*16	Conv2	1.3M	4.4M	x3.4
	1*32*32*3	16*3*3*3	1*32*32*16	Conv3	1.6M	6.5M	X4.1
	1*32*32*16	32*3*3*16	1*16*16*32	Conv4	2.8M	10.7M	x3.8
Depthwise Convolution	1*24*24*32	1*3*3*32	1*12*12*32	Depth1	676K	988K	x1.5
	1*48*48*8	1*3*3*8	1*48*48*8	Depth2	1.7M	3.6M	x2.1

a. Data Set b. Results

→ Kernel level speed-up

Application (CNN-based)		# Instructions		Reduction Factor
		NMC	Scalar	
Person Detection	MobileNet	58.9M	394M	x6.7
Image Classification	Resnet50	57.3M	608.9M	x10.6
Micro Speech	PureDepthNet	23M	32.7M	x1.4

Speed up **x10** for full CNN Resnet50

→ Application-level speed-up

Summary

- An NMC library was integrated in TFLM framework
  - Any Tiny ML model developed with TensorFlowLite can be run on the NMC architecture without any modification
- HW/SW co-design can be used to optimize NMC architectures
  - With a QEMU-based emulation architecture platform
  - To adapt NMC architecture to AI applications

Flexible Integration of a Neural Network Library in TensorFlow Lite Framework for Efficient Programmable Near-Memory Computing Architectures

T. Bricout, M. Kooli, J.-P. Noel, M. Ramirez Corrales, H.-P. Charles
Univ. Grenoble Alpes, CEA, LIST F-38000 Grenoble