

Importing and profiling a pre-trained machine learning model on PSOC™ Edge

Customer training workshop (CTW)

March 2026

Public

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 Machine learning: Development workflows 10

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

7 Lab 2: Performance improvement across CPUs and NPUs 25

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1	**Prerequisites**	**3**
2	PSOC™ Edge and ML development ecosystem	7
3	Machine learning: Development workflows	10
4	Using DEEPCRAFT™ Model Converter to deploy a pre-trained model	12
5	Hands-on labs	15
6	Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler	18
7	Lab 2: Performance improvement across CPUs and NPUs	25
8	Lab 3: Performance improvement with internal memories	31

Copyright © Infineon Technologies AG 2026. All rights reserved.

Prerequisites

- Recommended training prerequisites
  - Introduction to PSOC™ Edge and ModusToolbox™
    - For an introduction to PSOC™ MCUs, including a getting started guide to ModusToolbox™, visit:
      - https://www.infineon.com/product-information/psocdeveloper
    - For PSOC™ Edge trainings, from beginner tutorials to advanced trainings, visit:
      - PSOC™ Edge E84 Training Collection
  - Introduction to Edge AI and machine learning
    - Edge AI at Infineon.com
    - Introducing ML-enabled PSOC™ Edge for next generation Edge AI devices
    - DEEPCRAFT™ AI Hub



Copyright © Infineon Technologies AG 2026. All rights reserved.

Prerequisites: Tools

- Development tools for all labs (see training manual for links)
  - ModusToolbox™ v3.7
  - Eclipse IDE for ModusToolbox™ v2025.8.0
  - Edge Protect Security Suite v1.6
  - ModusToolbox™ Programming Tools v1.7.0
  - DEEPCRAFT™ Model Converter v5.8.4154
  - ModusToolbox™ Machine Learning Pack v3.0.1
  - KIT_PSE84_EVAL_EPC2 BSP v1.0.0
  - Terminal Emulator
  - KIT_PSE84_EVAL



* Other tool versions and hardware can be used with minimal or no modifications to the training manual.

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 **PSOC™ Edge and ML development ecosystem** **7**

3 Machine learning: Development workflows 10

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

7 Lab 2: Performance improvement across CPUs and NPUs 25

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

PSOC™ Edge and ML development ecosystem

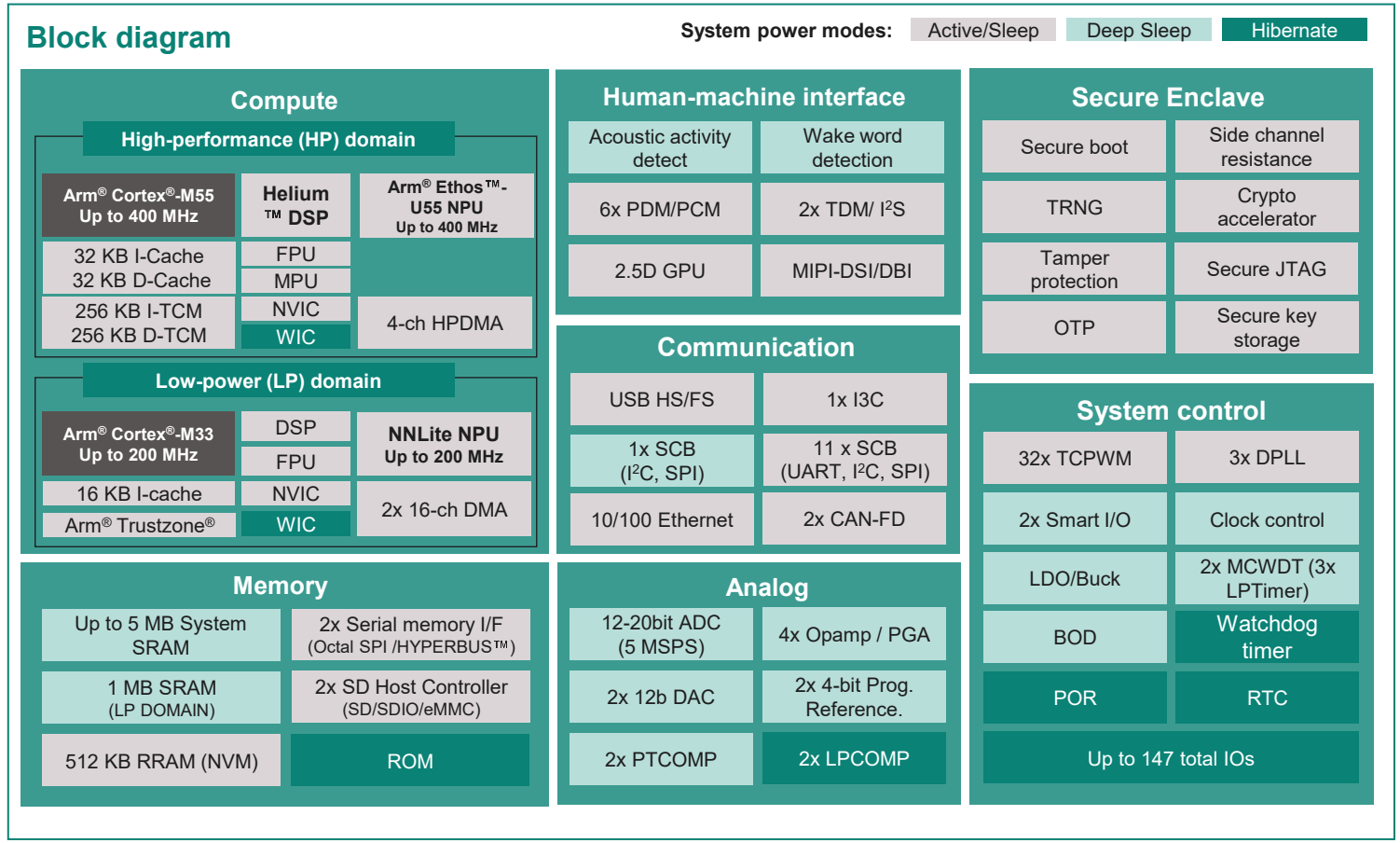
PSOC™ Edge E84: Low-power advanced ML MCU with graphics

Product highlights

- **High-performance, real-time compute domain:**
  - **Cortex®-M55 with Helium™ DSP and Ethos™-U55 NPU for ML**
  - **Up to 5 MB system SRAM**, 256 KB I&D TCMs
- **Low-power, real-time compute domain:**
  - **Cortex®-M33 and DSP with Infineon's NNLite for ML**
  - **512 KB RRAM, 1 MB SRAM**
- **HMI:**
  - Traditional MCU HMI
  - Local voice, cloud voice
  - Vision for friction-free interface & safety
  - Low-power graphics, up to 1024x768, and MIPI-DSI/DBI
  - **ML: Advanced ML leveraging Ethos™-U55 and NNLite**
- **Peripherals and IO:**
  - USB, 10/100 Ethernet, CAN, SPI, UART, I2C, I3C, and I2S
  - Ultra-low-power always-on analog
- **Security:**
  - Secure Enclave at 25 JIL pts, Infineon Edge Protect Category 4 (EPC4) with PSA certified L3/L4 iSE
  - Edge Protect Bootloader and TF-M stack available in ModusToolbox™
- **Operating range:**
  - – 20 to 70°C Ta (consumer), –40 to 105°C Ta (industrial)

Block diagram

System power modes: Active/Sleep | Deep Sleep | Hibernate



Status/availability

- **Documentation** available now
- **Samples and evaluation kits** available (PSOC™ Edge E84)
- **Qualified samples** available now

Packages

WLB154 (5.2 x 4.3 x 0.5 mm, 0.35 mm)

BGA220 (10 x 10 x 1.2 mm, 0.65 mm)

eWLB235 (7 x 7 x 0.7 mm, 0.4 mm)

Operating range

–20 to 70°C Ta (Consumer), –40 to 105°C Ta (Industrial)

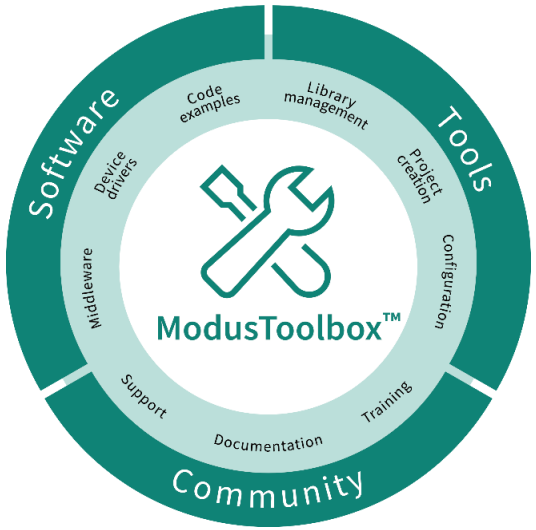
Applications and target markets

- Smart home, appliances, residential AC, wearables, industrial HMI, smart speakers, and more

Copyright © Infineon Technologies AG 2026. All rights reserved.

PSOC™ Edge – Ecosystem for AI/ML applications

ModusToolbox™



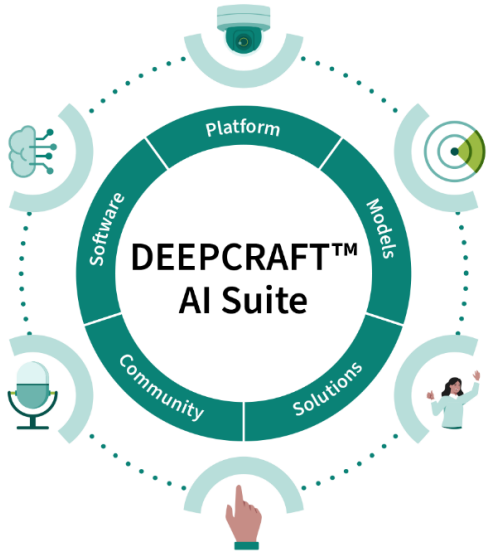
ModusToolbox™ – Comprehensive collection of resources

- Flexible development tools and resources
- Robust software, middleware, and security ecosystem

Zephyr

Comprehensive Zephyr RTOS enablement for seamless integration and rapid development

DEEPCRAFT™ AI Suite



Accelerate Edge AI Innovation with DEEPCRAFT™ AI Suite

- Integrated platform and tools for end-to-end AI workflow
- Comprehensive AI solutions
- Centralized DEEPCRAFT™ AI Hub

PSOC™ Edge embedded development





Development boards

- **PSOC™ Edge E84 Evaluation Kit:** PSOC™ Edge device functional evaluation
  - **PSOC™ Edge E84 SOM:** PSOC™ Edge Minimal system-on-module
- **PSOC™ Edge E84 AI Kit:** Low-cost kit with multiple sensors for evaluation of AI capabilities and fast prototyping

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 **Machine learning: Development workflows** **10**

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

7 Lab 2: Performance improvement across CPUs and NPUs 25

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

Machine learning: Development workflows

Collect or Bring your own Data and Build a Custom Model



DEEPCRAFT™ Studio + ModusToolbox™

Bring your own Model!



DEEPCRAFT™ Model Converter + ModusToolbox™

Directly add ML Capability or Ready-to-use Solution



DEEPCRAFT™ Ready Models and Solutions

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 Machine learning: Development workflows 10

4 **Using DEEPCRAFT™ Model Converter to deploy a pre-trained model** **12**

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

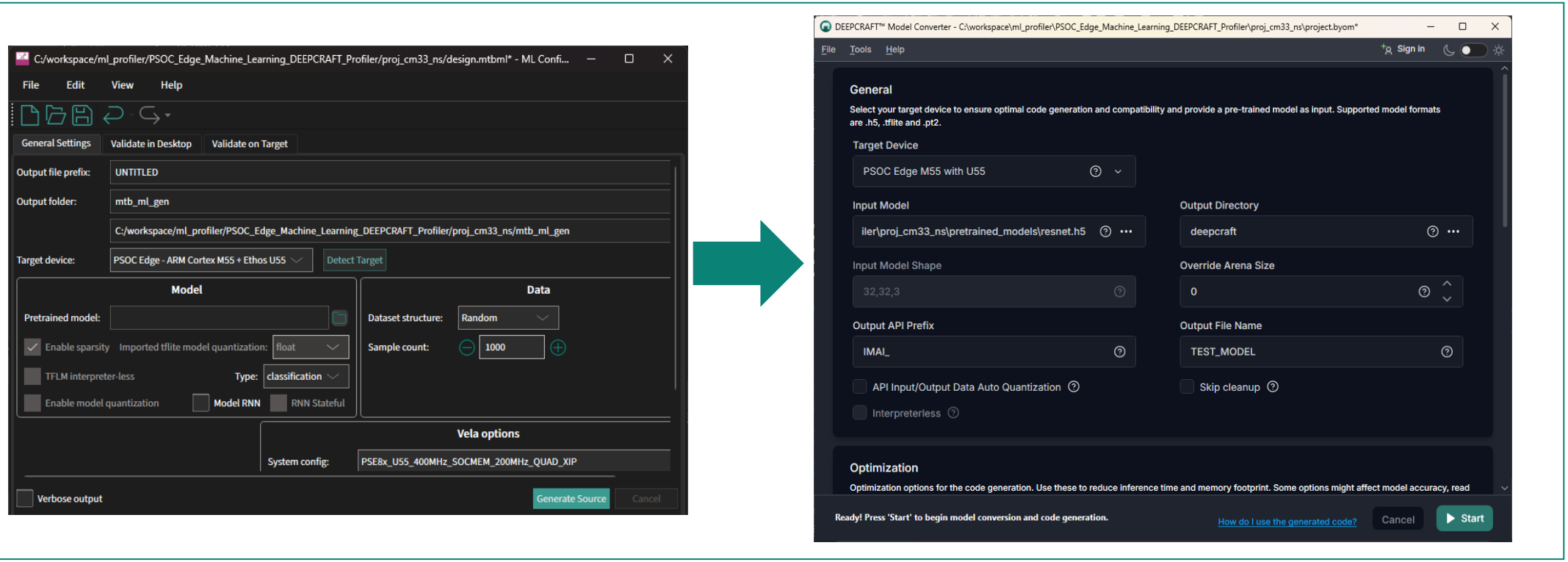
7 Lab 2: Performance improvement across CPUs and NPUs 25

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

DEEPCRAFT™ Model Converter

- The DEEPCRAFT™ Model Converter is replacing the ModusToolbox™ machine learning (MTBML) GUI, but there are some limitations
  - Desktop validation
  - Validation data streaming
  - TFLM interpreter-less



Copyright © Infineon Technologies AG 2026. All rights reserved.

DEEPCRAFT™ model profiling enablement

Model source generated with DEEPCRAFT™ Studio or Model Converter support profiling

Required	Located in DEEPCRAFT™ Profiler CE	Description
IMAI_PROFILING	Makefile(cm33/cm55)	Defined in the project Makefile to configure DEEPCRAFT™ source for profiling
IMAI_hook_region function	ml_validation.c	User defined function used by DEEPCRAFT™ source to keep track of timing
Elapsed timer	elapsed_timer.c/h	User defined functions which return elapsed time, uses systick

- **IMAI_init**: Initialises the model
- **IMAI_compute**: Pass in validation data to the model, results returned in a buffer
- **IMAI_mtb_models_profile_log**: Outputs CPU/NPU cycle time per inference
- **IMAI_mtb_models_print_info**: Outputs model memory/tensor size

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1	Prerequisites	3
2	PSOC™ Edge and ML development ecosystem	7
3	Machine learning: Development workflows	10
4	Using DEEPCRAFT™ Model Converter to deploy a pre-trained model	12
5	**Hands-on labs**	**15**
6	Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler	18
7	Lab 2: Performance improvement across CPUs and NPUs	25
8	Lab 3: Performance improvement with internal memories	31

Copyright © Infineon Technologies AG 2026. All rights reserved.

Hands-on labs

Lab No.	Test case	Model	Core(s) tested	User task
1	Generate model source code on Model Converter and profile it on device	Resnet int8 and float models	CM33 + NNLite	Compare memory and accuracy. Make note of the M33+NNLite CPU and NPU cycles
2	Performance Improvements across cores	Resnet int8 model	CM33 + NNLite CM33 CM55 + U55	Fill in cycle count table
3	Memory vs performance improvement: Internal memory case	Resnet int8 model	CM55 + U55: Model in SystemRAM Arena in SystemRAM	Fill in cycle count table
	Memory vs performance improvement: Flash case	Resnet int8 model	CM55 + U55: Model in ext. flash Arena in SystemRAM	Fill in cycle count table

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 Machine learning: Development workflows 10

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 **Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18**

7 Lab 2: Performance improvement across CPUs and NPUs 25

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

Lab 1: Creating ML DEEPCRAFT™ profiler code example

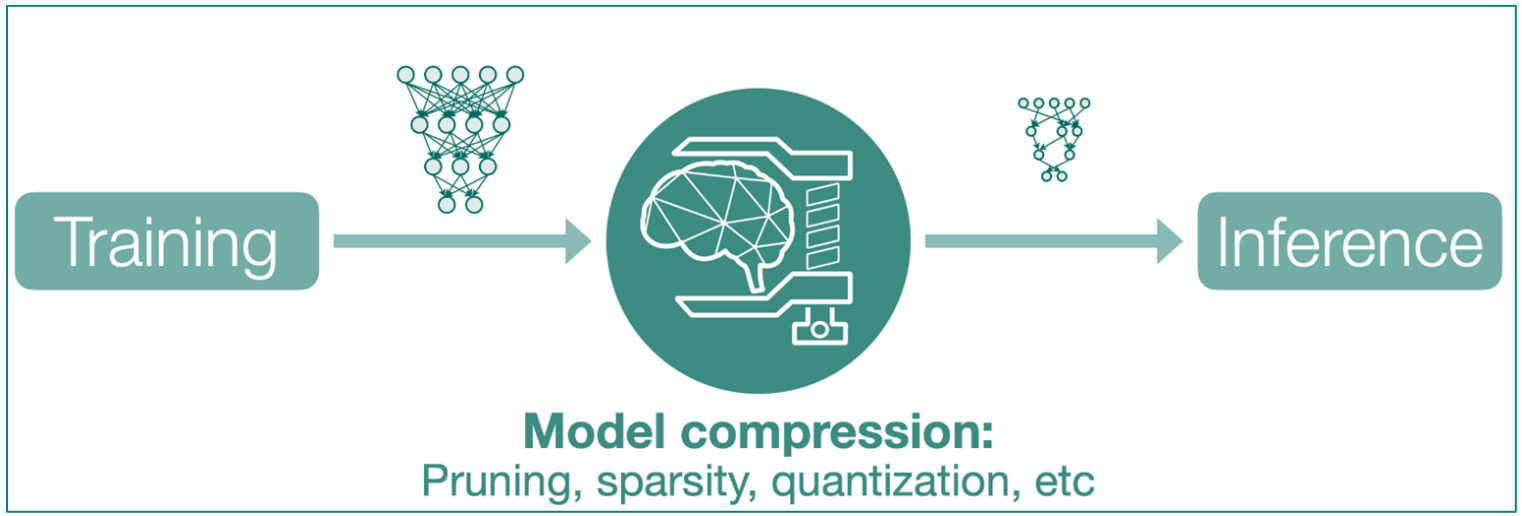
What will you learn?

- How to deploy a pre-trained model on PSOC™ Edge
- Understand the ML DEEPCRAFT™ profiler code example
- How to use this code example to profile NN models

Optimizing the NN model

- Why we need to optimize the NN models
  - Optimizes for size requiring lower memories on the edge device
  - Reduced power consumption
  - Improves computational efficiency

- NN models optimization techniques
  - Pruning
  - Sparsity
  - Quantization



Copyright © Infineon Technologies AG 2026. All rights reserved.

Configurations supported in DEEPCRAFT™ Model Converter

Model Converter configuration	Description
Target device	Currently CM33+NNLite and CM55+U55 targets are supported
Optimization: Sparsity	These optimizations are used to reduce storage and computations required for a model
Optimization: Network quantization	Quantization is used to reduce the precision of numerical data in a model, thereby decreasing the memory and computational requirements

Copyright © Infineon Technologies AG 2026. All rights reserved.

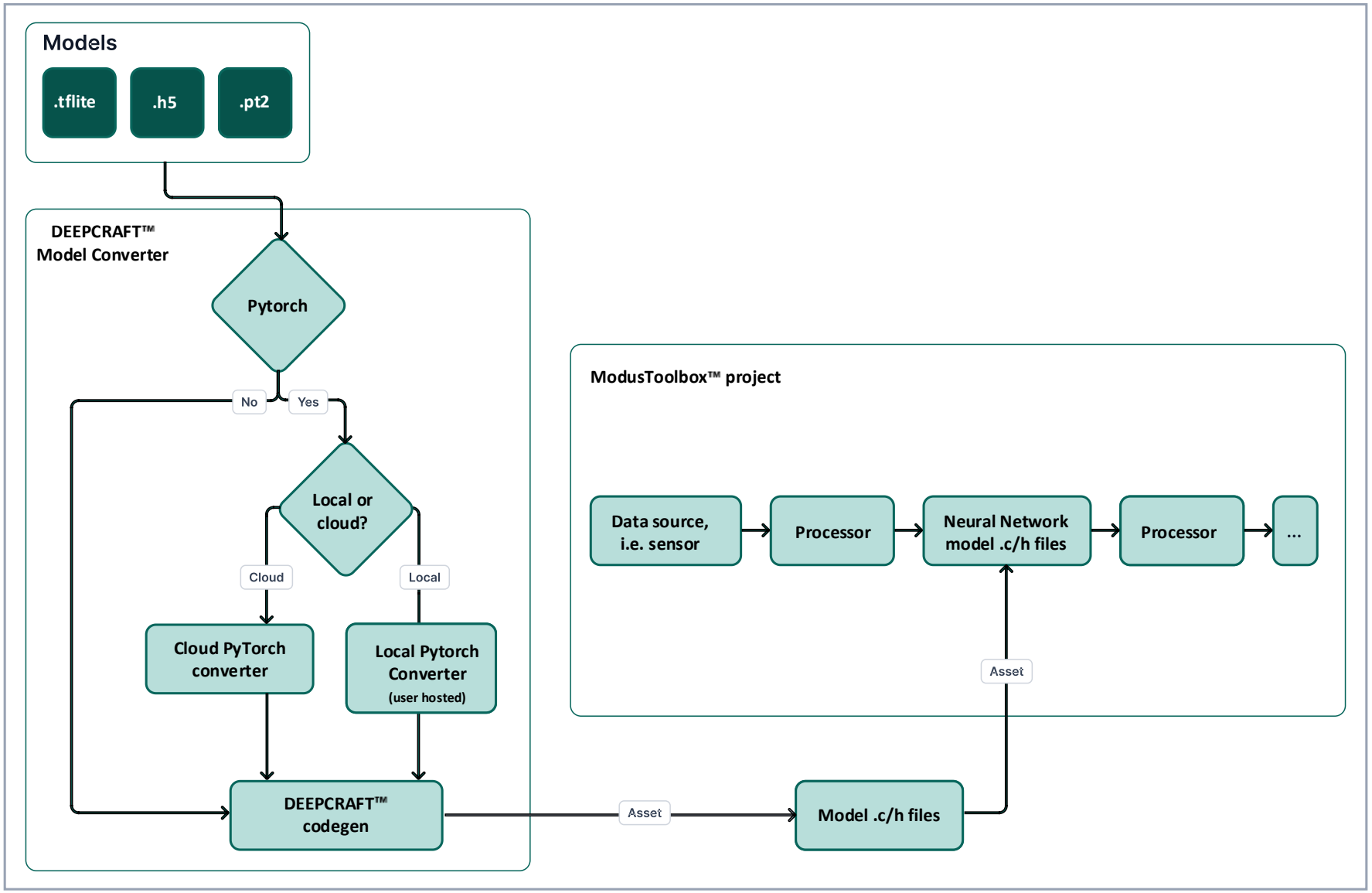
Model optimization techniques: Quantization

- **Quantization**
  - Keras (.h5) models support Int8×8 or Int16×8 by providing representative input samples for calibration
  - PyTorch (.pt2) models can be quantized to Int8x8 quantization only
  - The input and output must use the same data type
- Model formats based on the quantization option in Model Converter

Supported quantization	TensorFlow Keras (.h5)	TensorFlow Lite (.tflite)	PyTorch (.pt2)
Float 32 activations + Float 32 weights	✓	✓	✓
Int8 activations + Int8 weights	✓	✓	✓
Int16 activations + Int8 weights	✓	✓	x

Copyright © Infineon Technologies AG 2026. All rights reserved.

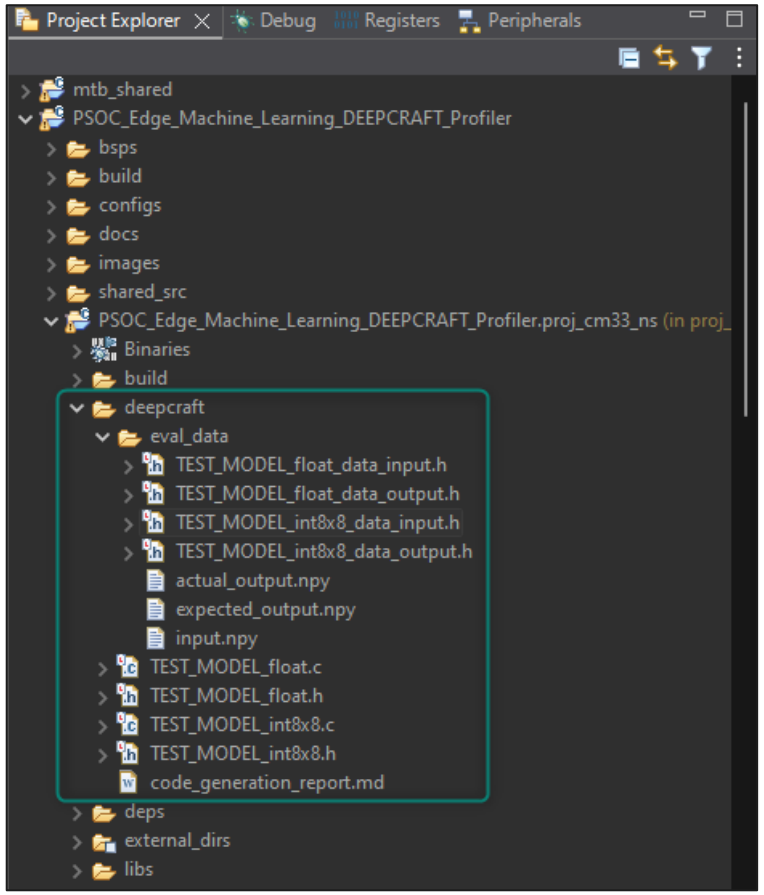
Generate model source and deploy on Edge



Copyright © Infineon Technologies AG 2026. All rights reserved.

Results

Generated model source for the M33 by importing a model



Compare estimated int8x8 and float memory/accuracy

Data Type	Model Memory, bytes	Tensor Arena, bytes	CPU Cycles	NPU Cycles
float	316992	217088	Unknown	Unknown

```
Classification Metrics:
(All results are relative to the original model)
Argmax Errors:          0 / 100 samples
Argmax Similarity:      100.00%
```

Data Type	Model Memory, bytes	Tensor Arena, bytes	CPU Cycles	NPU Cycles
int8x8	94088	59392	Unknown	Unknown

```
Classification Metrics:
(All results are relative to the original model)
Argmax Errors:          1 / 100 samples
Argmax Similarity:      99.00%
```

Compare estimated vs actual int8x8 memory/accuracy

```
****************** PSOC Edge MCU: Machine Learning DEEPCRAFT Profiler on CM33+NNLite ******************

Loaded 1 models:
Model 0:

**************************************************
MTB Tflite-Micro inference
Model name              :       resnet_int8x8
Model size              :       91.88   kB
Buffer size             :       52.02   kB


Profiling 1 models:
Model 0: resnet_int8x8:
PROFILE_INFO, MTB ML model profile, avg_cpu_cyc=204190.12 , peak_cpu_cyc=217329, peak_cpu_frame=0, cpu_freq_Mhz=200
PROFILE_INFO, MTB ML model profile, avg_npu_cyc=5140238.50, peak_npu_cyc=5143065, peak_npu_frame=0, npu_freq_Mhz=200

**************************************************
FAIL with accuracy percentage =97.00, total_cnt=100
**************************************************

Profiling completed!
```


Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 Machine learning: Development workflows 10

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

7 **Lab 2: Performance improvement across CPUs and NPUs** **25**

8 Lab 3: Performance improvement with internal memories 31

Copyright © Infineon Technologies AG 2026. All rights reserved.

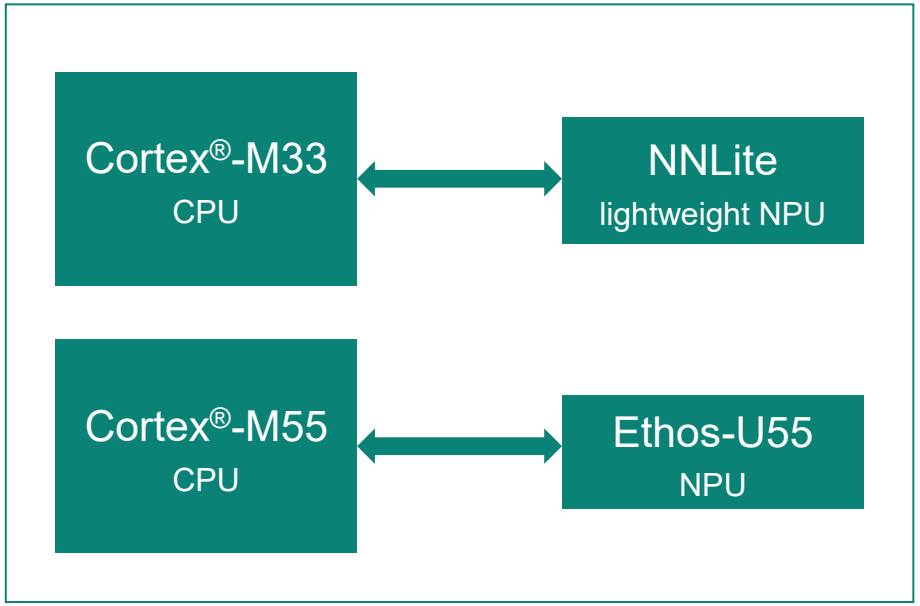
Lab 2: Performance improvement across CPUs and NPUs

What will you learn?

- The user will learn about the use of the Vela compiler to accelerate the NN execution with NPU
- Performance improvement with NPUs by comparing the cycle count of the CM33, CM33+NNLite, and CM55 + U55

PSOC™ Edge NPUs

- Infineon NNLite NPU on Cortex®-M33
  - Accelerates neural networks for ML and AI applications
  - Optimises key kernel functions: Dense, Convolution, and Pooling
  - Efficiently supports activations: ReLU, tanh, Leaky ReLU, and Sigmoid
  - Low-power consumption: Four simultaneous 8-bit MACs per cycle
- Ethos-U55 NPU on Cortex®-M55
  - Offloads intensive ML computations from CM55
  - Enhances performance and energy efficiency
  - Supports SIMD operations with Helium technology
  - Ideal for image and speech recognition
- Advanced ML capabilities
  - Supports complex kernels: LSTM, GRU
  - Suitable for Wake word detection (WWD) and Human activity recognition (HAR)
  - Configurable NN topologies with single-layer acceleration



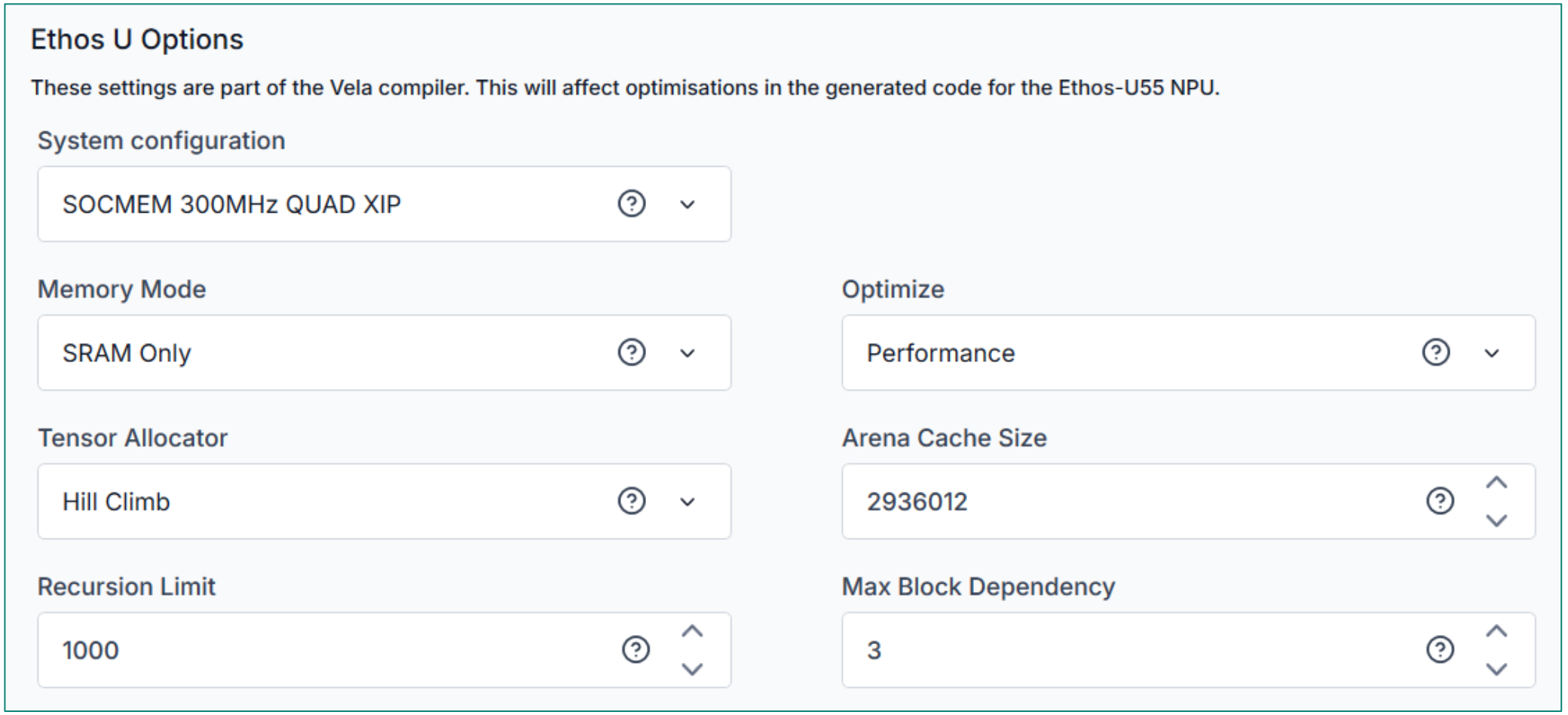
Copyright © Infineon Technologies AG 2026. All rights reserved.

NN kernels supported by NPUs

NNLite 4 MAC	U55 128 MAC
Kernels/Operators supported – Pointwise: MUL, SUB – Convolution: CONV_2D, DEPTHWISE_CONV_2D, AVERAGE_POOL_2D, and MAX_POOL_2D – Dense: FULLY_CONNECTED – RNN: UNIDIRECTIONAL_SEQUENCE_LSTM **Activation functions supported** – TANH – LEAKY_RELU – LOGISTIC – RELU	**Kernels/Operators supported** – Pointwise: ADD, MUL, and SUB – Convolution: CONV_2D, DEPTHWISE_CONV_2D, AVERAGE_POOL_2D, and MAX_POOL_2D – Dense: FULLY_CONNECTED – RNN: UNIDIRECTIONAL_SEQUENCE_LSTM **Activation functions supported** – TANH – LEAKY_RELU – LOGISTIC – RELU

Copyright © Infineon Technologies AG 2026. All rights reserved.

Vela compiler for U55 NPU



Copyright © Infineon Technologies AG 2026. All rights reserved.

Performance improvement across PSOC™ Edge cores

Fill table at the end of this Lab:

Memory	Average cycle count	Peak cycle count
CM33 Only	–	–
CM33+NNLite	–	–
CM55 + U55	–	–

Copyright © Infineon Technologies AG 2026. All rights reserved.

Lab 2: Results

Expected results

Memory	Average cycle count	Peak cycle count
CM33 Only	CM33: 42,191,164	CM33: 42,197,220
CM33+NNLite	CM33: 204,190 NNLite: 5,140,238	CM33: 217,329 NNLite: 5,143,065
CM55 + U55	CM55: 88,344 U55: 183,088	CM55: 120,287 U55: 186,437

Inference times

- CM33: 211 ms at 200 MHz
- CM33+NNLite: 26.7 ms at 200 MHz
- CM55+U55: 0.68 ms at 400 MHz

Copyright © Infineon Technologies AG 2026. All rights reserved.

Table of contents

1 Prerequisites 3

2 PSOC™ Edge and ML development ecosystem 7

3 Machine learning: Development workflows 10

4 Using DEEPCRAFT™ Model Converter to deploy a pre-trained model 12

5 Hands-on labs 15

6 Lab 1: Deploy a model on PSOC™ Edge using the Machine Learning DEEPCRAFT™ profiler 18

7 Lab 2: Performance improvement across CPUs and NPUs 25

8 **Lab 3: Performance improvement with internal memories** **31**

Copyright © Infineon Technologies AG 2026. All rights reserved.

Lab 3: Performance improvement with internal memories

What will you learn?

- Details of memories available in PSOC™ Edge MCU
- Performance improvement with internal memories

Lab 3: Performance improvement with internal memories

Memory	Size	Power Domain	Description
SRAM	1MB	PD0	SRAM memory is used for storing data, variables, and buffers required for runtime operations. The MCU can read, write, and manipulate data during program execution
RRAM	Up to 512 KB	PD0	RRAM is typically used to store CPU instructions and data when the device power is off. It offers high-bulk program performance and supports ultra-low-power operation
System SRAM (SoCMEM)	Up to 5120 KB	PD1	System SRAM in the PD1 domain is used for storing data, variables, and buffers required for runtime operations
ITCM	256 KB	PD1	This can be utilized for implementing essential routines like interrupt handling or real-time tasks, guaranteeing maximum responsiveness without any cache latency
DTCM	256 KB	PD1	To achieve faster, predictable, and deterministic data access, the CPU core is directly linked to DTCM for storing critical buffers and efficiently managing context-switching data in systems with rapid context switching
SMIF Core 0/Core 1	64 MB/64 MB in XIP	PD1	Extend the memory used for code or data by 1GB per SMIF, or 64 MB per SMIF core when used in XIP mode

Copyright © Infineon Technologies AG 2026. All rights reserved.

Memories vs ML performance of PSOC™ edge

Fill table at the end of this Lab:

Core	Model on system SRAM Tensor Arena on system SRAM		Model on external flash Tensor Arena on system SRAM	
	Average cycle count	Peak cycle count	Average cycle count	Peak cycle count
CM55 + U55	–	–	–	–

Copyright © Infineon Technologies AG 2026. All rights reserved.

Results

Expected results

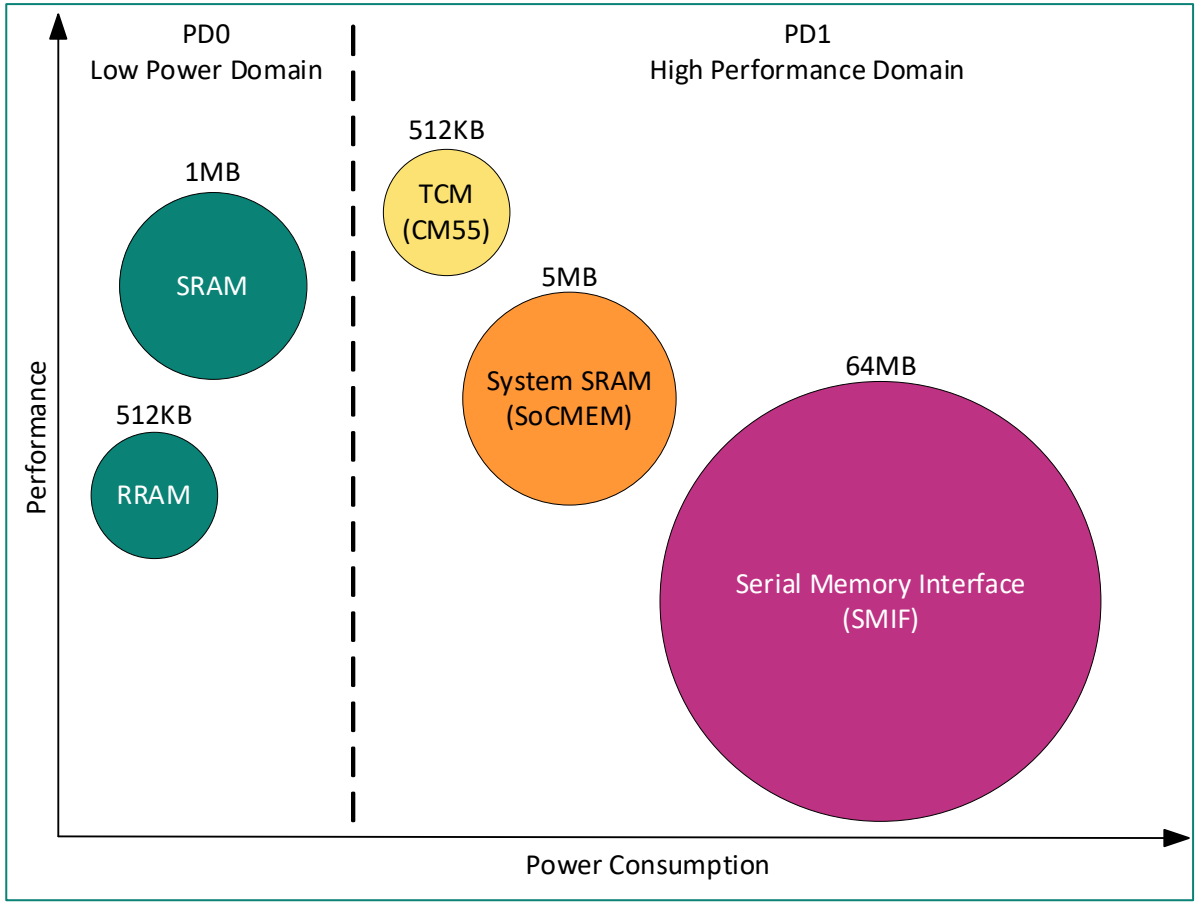
Core	Model on system SRAM Tensor Arena on system SRAM		Model on Ext. Flash Tensor Arena on system SRAM	
	Average cycle count	**Peak cycle count**	**Average cycle count**	**Peak cycle count**
CM55 + U55	CM55: 88, 344 U55: 183, 088	CM55: 120, 287 U55: 186, 487	CM55: 92,116 U55: 732, 748	CM55: 131, 000 U55: 738, 464

Optimizing model and tensor arena by placing them in System SRAM reduced the inference time by ~3x.

Copyright © Infineon Technologies AG 2026. All rights reserved.

Power and performance impact for accessing memories

- RRAM or SRAM is preferred for CM33/CM33+NNLite use case
- System SRAM is preferred for CM55/CM55+U55 use cases



Copyright © Infineon Technologies AG 2026. All rights reserved.

Revision history

Document revision	Date	Description of change
**	2026-03-15	Initial release.

Copyright © Infineon Technologies AG 2026. All rights reserved.


Infineon