

tinyML[®] EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event



www.tinyML.org



VOICE YOUR FUTURE

DSP®
GROUP

DSP Group

Innovative Minimization of Parameter
Memory in small silicon low power
devices

June 2021



Moshe Haiut is a principle senior engineer in the CTO team at DSP Group, Herzliya, Israel. His main areas of expertise are communication, video and audio digital signal processing, digital hardware architecture, and neural networks. Moshe holds M.Sc (with honors) and B.Sc. (with honors) degrees in Electrical Engineering from Tel-Aviv University and the Technion, Israel. Moshe is the architect of the nNetLite – an ultra-low-power H/W engine for NN inference models execution

The Problem – Memory space for the parameters (weights)

- Small-to-medium NN models may have millions of parameters, while the tinyML SOC usually has limited memory (hundreds of KBytes) to store the weights data
- In this presentation we show how this problem was solved by an innovative h/w accelerator (Weights Extraction Unit) combined with a powerful s/w toolchain

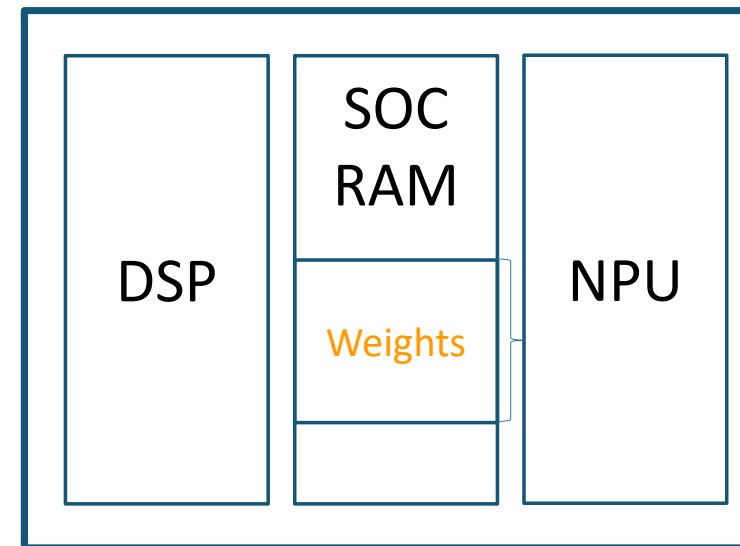


Quantization

Pruning

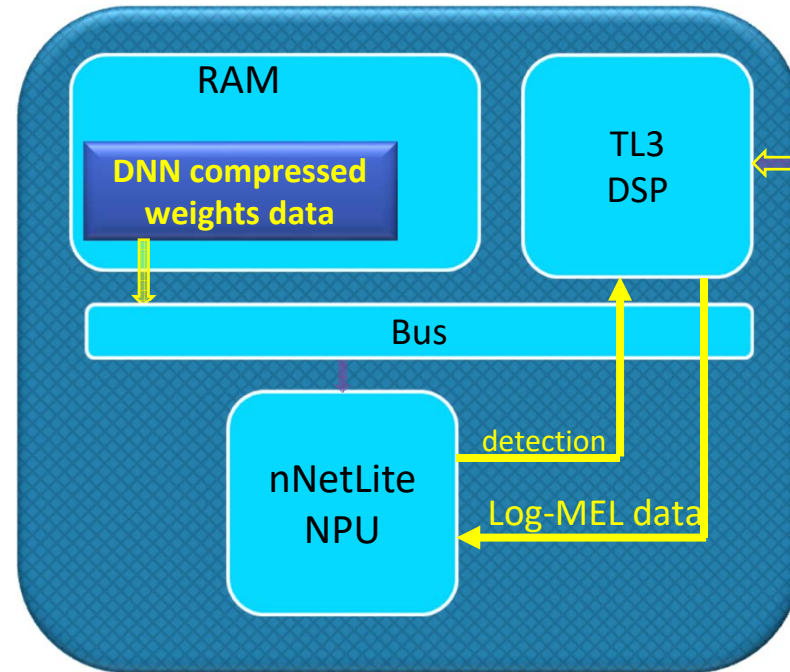
Entropy Coding

Packing



The DSPG nNetLite h/w Engine

- The nNetLite engine is a stand-alone module, planned to be embedded in DSP Group's future SOCs.
- The DBM10L ultra-low-power device comprises the nNetLite engine and the CEVA TeakLite-3 side-by-side



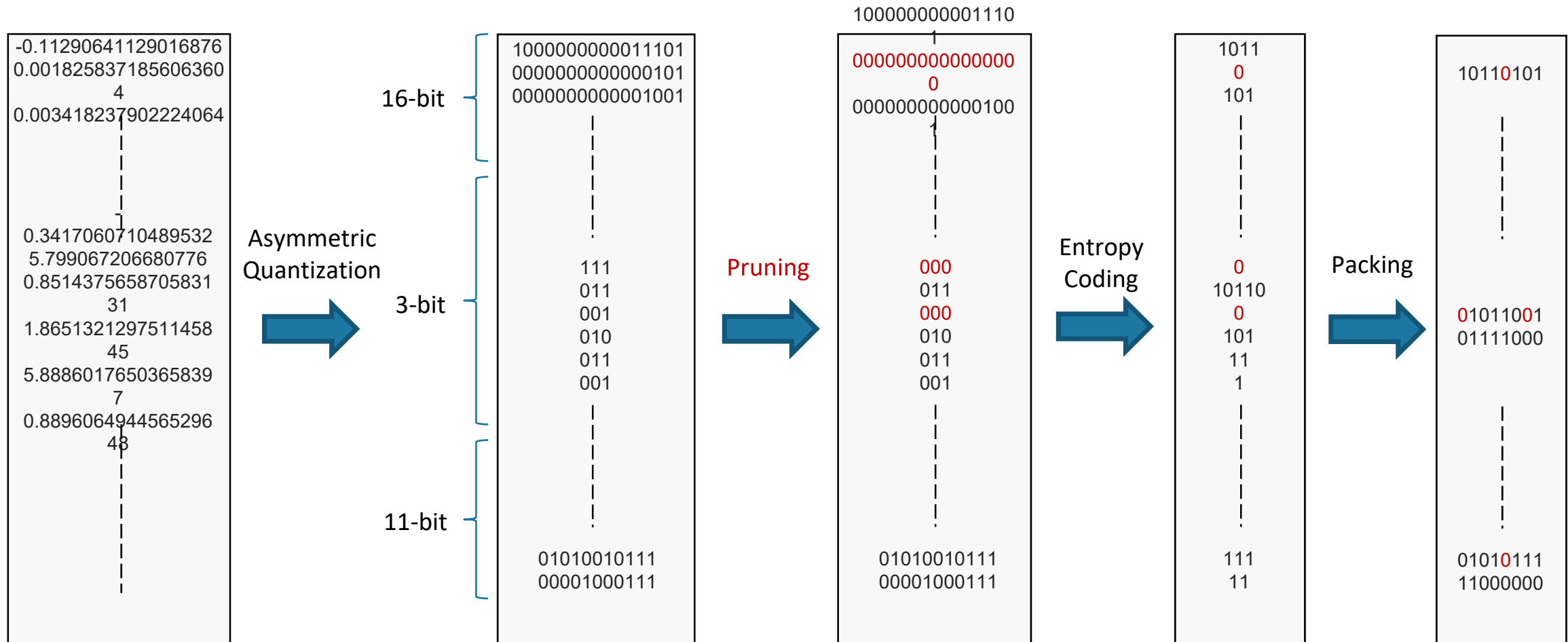
DBM10L incorporates the nNet Lite engine and a DSP

The nNet Lite engine is configured by the DSP and executes concurrent inferences in a stand-alone manner

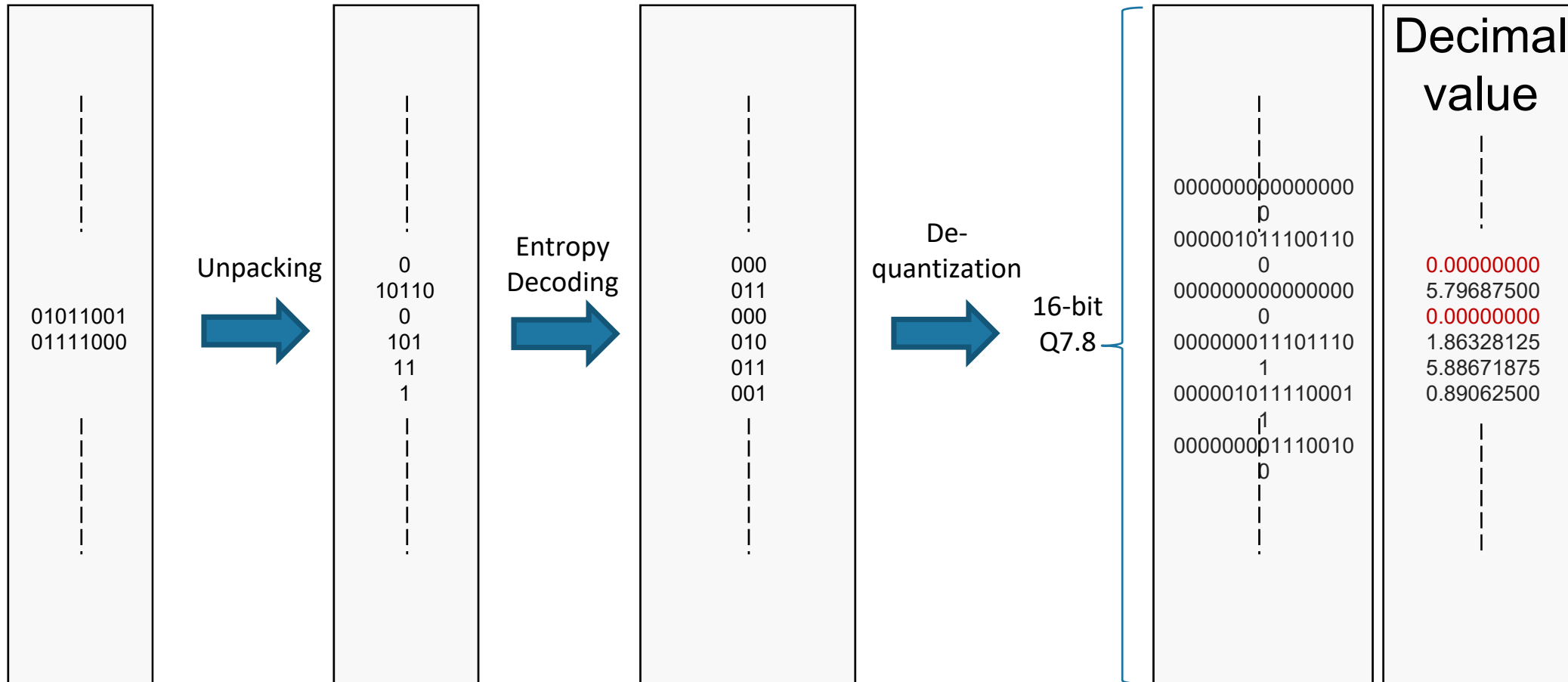
The DSP makes the audio pre-processing (e.g. log-MEL calc.) and the detection post processing to support the concurrent inferences

The NN weights are stored in the shared RAM in a compressed form and fetched by the nNet Lite engine via DMA

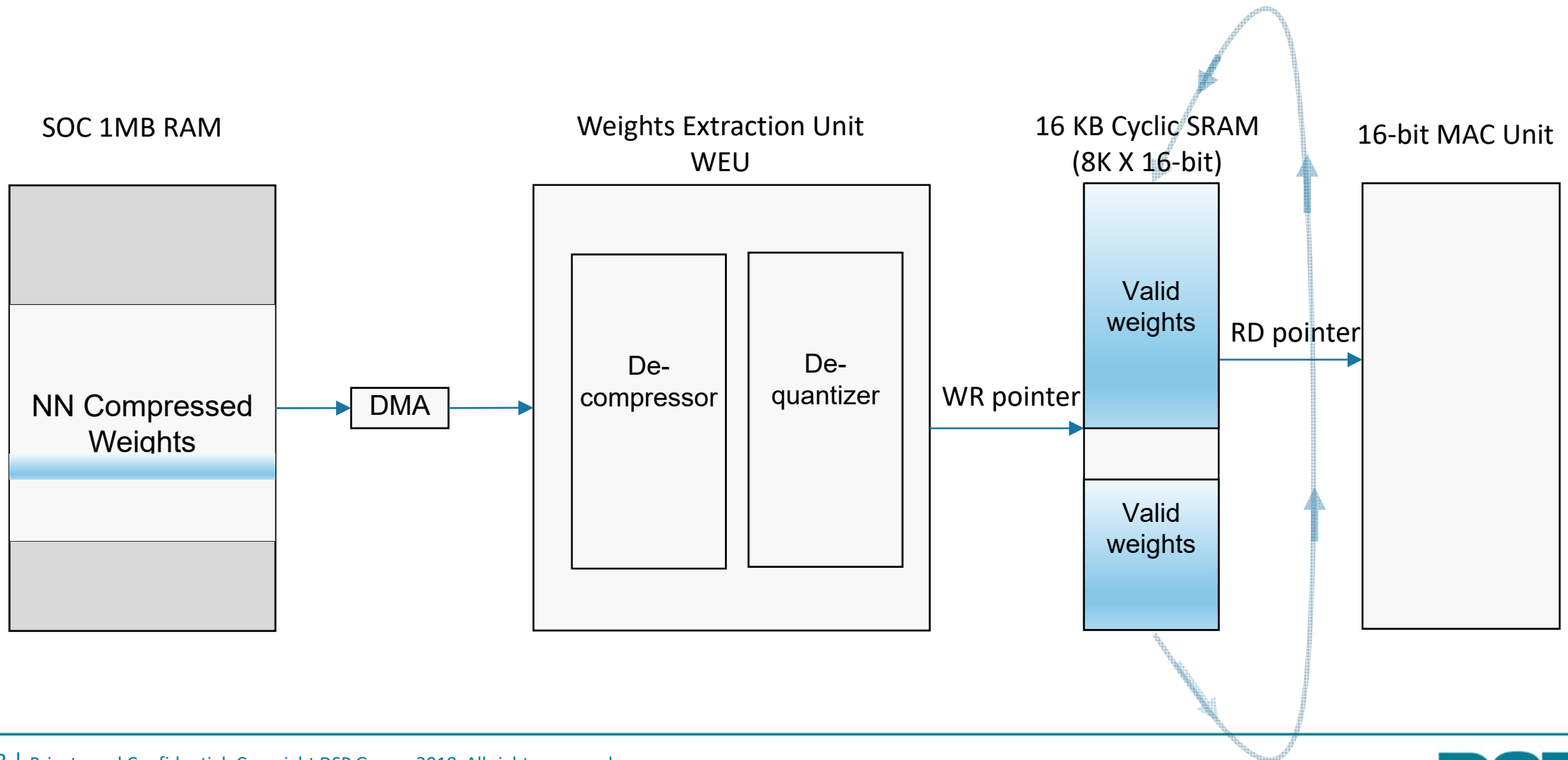
Weights Compression by the nNetLite Compiler



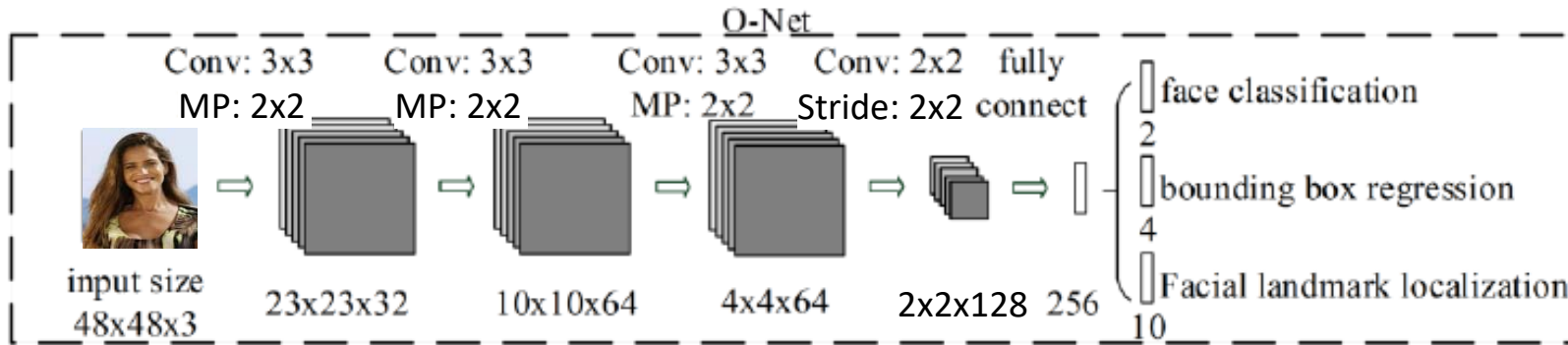
Weights De-compression by the nNetLite WEU h/w module



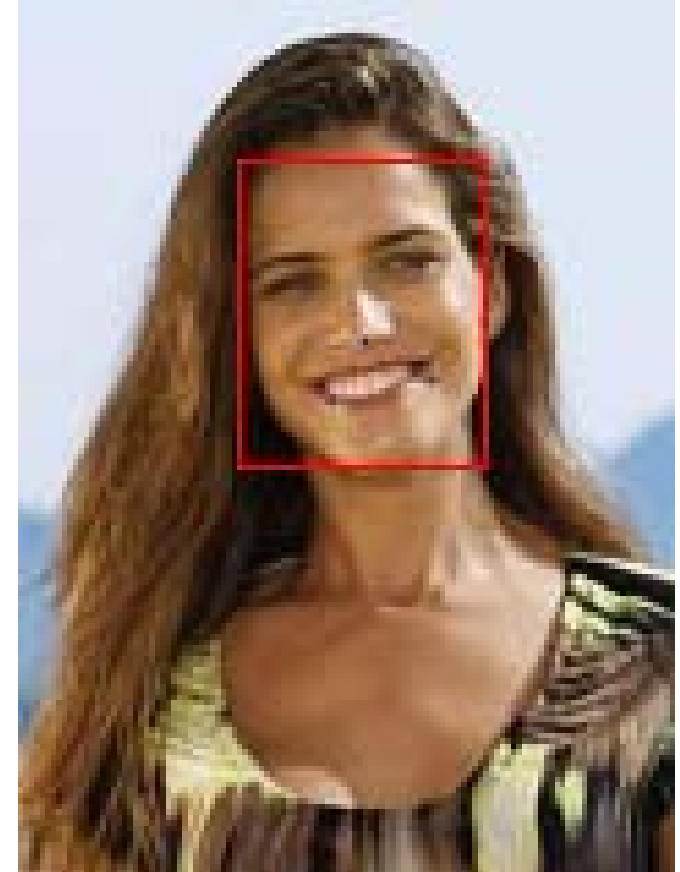
Weights De-compression by the nNetLite WEU h/w module



Example: Face Detection NN



- Number of layers: 8
- Number of MAC ops: ~12M / Inference
- Number of Weights: 224,656
- Available memory: 280KB



w/o Compression. Weights: 439 KB, RMSE: 0.130191

DSP Group's NPU SDK --- Version 1.20.0

C:/Users/mosheh/Documents/WORK/nNet Lite/Imaging NN for nNet Lite/Ver.Feb_14_2021/Cmpressed/O-NET/TEMP/Onet.onnx

Open model → Parameters → **Weights quantization** → Final layers setup → Weights compression

	Number of weights	Resolution	Weight mode	bias	shift	gain	gain offset
0	896	16	16-bit Q7.8	-252	0	7	0
1	10106	16	16-bit Q7.8	-1740	0	8	0
2	36928	16	16-bit Q7.8	-1508	0	8	0
3	32896	16	16-bit Q7.8	-709	0	8	0
4	131328	16	16-bit Q7.8	-1982	0	8	0
5	10106	16	16-bit Q7.8	2445	0	8	0
6	514	16	16-bit Q7.8	-158	0	7	0
7	2570	15	16-bit Q7.8	6578	0	7	0

Output weights file: 448864 bytes
Number of parameters: 224656

Back Next

nNetLite Compiler

100% 10/10 (accuracy 0%)

Power Consumption

Calculate

Rate: 1 Inf/sec | NPU Clock: 12 MHz | DBM10L Power: 566.48 uW

Logging and Analysis

Layer logging settings

Layers to Log: 7: | ECU Operations | Select All | Reset Selection | Show Analysis

Analysis

Results Source: File Folder | 14_2021/Cmpressed/O-NET/TEMP/test_vectors/sim_layers_log_20210510_125540 | Set Results

Layer: 7 | Layer Type: Label Prediction | SoftMax | Analyze

Status

Vectors Correlation:

- MSE (Mean Square Error): 0.016943
- RMSE (Root Mean Square Error): **0.130191**
- NRMSE1 (Normalized RMSE [RMSE / Mean(Ref)]): 0.258395
- NRMSE2 (Normalized RMSE [RMSE / Range(Ref)]): 0.273477
- ME (Mean Error): -0.0634533
- AME (Absolute Mean Error): 0.104741
- SDR (Standard Deviation of Residuals): 0.113681
- RAE (Relative Absolute Error): 1.04542
- RRSE (Root Relative Squared Error [RMSE / Stdev(ref)]): 1.11947
- MAPE (Mean Absolute Percentage Error): 0.282729

Elapsed Time: 00:00:48 | Clear

nNetLite Simulator

File Explorer: Imaging NN for nNet Lite > Ver.Feb_14_2021 > Cmpressed > O-NET > TEMP

Name	Date modified	Type	Size
ONet_task_batch.json	10/05/2021 12:48	JSON File	1 KB
ONet_task_batch.lst	10/05/2021 12:48	MASM Listing	1 KB
ONet_weights.dat	10/05/2021 12:48	DAT File	439 KB

weights binary size

12-bit weights, 50% Prun. Weights: 273 KB, RMSE: 0.123725

DSP Group's NPU SDK --- Version 1.20.0

C:/Users/mosheh/Documents/WORK/nNet Lite/Imaging NN for nNet Lite/Ver.Feb_14_2021/Cmpressed/O-NET/TEMP/ONet.onnx

Open model → Parameters → Weights quantization → Final layers setup → Weights compression

	Number of weights	Resolution	Weight mode	bias	shift	gain
0	896	16	16-bit Q7.8	-252	0	
1	18406	16	16-bit Q7.8	-1740	0	
2	36928	12	16-bit Q7.8	-1508	4	
3	32896	12	16-bit Q7.8	-709	4	
4	131328	12	16-bit Q7.8	-1982	4	
5	1628	15	16-bit Q7.8	2445	0	
6	514	16	16-bit Q7.8	-158	0	
7	2570	15	16-bit Q7.8	6578	0	

Output weights: 342,208 bytes
Number of parameters: 224,116

Open model → Parameters → Compress weights
 No Pruning

% To Prune

0	5
1	5
2	50
3	50
4	50
5	5
6	5
7	5

nNetLite Compiler

File Explorer: WORK > nNet Lite > Imaging NN for nNet Lite > Ver.Feb_14_2021 > Cmpressed > O-NET > TEMP

Name	Date modified	Type	Size
ONet.onnx	14/02/2021 10:42	ONNX File	
ONet_compressed_weights.dat	10/05/2021 13:20	DAT File	273 KB

weights binary size

100% 10/10 (accuracy 0%)

Power Consumption

Calculate

Rate	NPU Clock	DBM10L Power
1 Inf/sec	12 MHz	565.912 uW

Logging and Analysis

Layer logging settings

Layers to Log: 7: ECU Operations Select All Reset Selection Show Analysis

Analysis

Results Source: Folder: _14_2021/Cmpressed/O-NET/TEMP/test_vectors/sim_layers_log_20210510_132604 Set Results

Layer: 7 Layer Type: Label Prediction SoftMax Analyze

Status

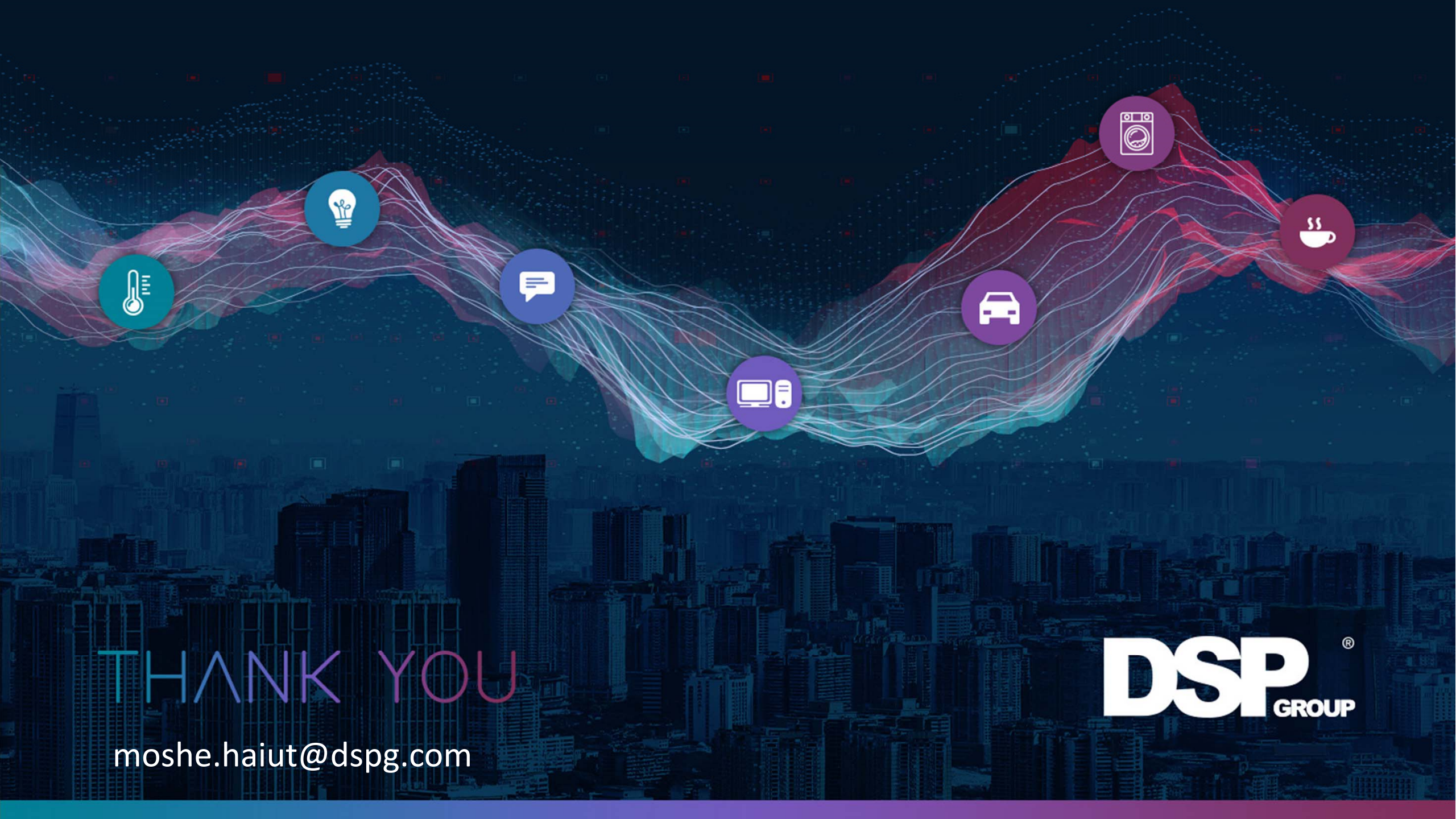
Variance: 0.0129137
Stdev: 0.113638

Vectors Correlation:

MSE (Mean Square Error):	0.0153078
RMSE (Root Mean Square Error):	0.123725
NRMSE1 (Normalized RMSE [RMSE / Mean(Ref)]):	0.245563
NRMSE2 (Normalized RMSE [RMSE / Range(Ref)]):	0.259895
ME (Mean Error):	-0.0604846
AME (Absolute Mean Error):	0.0999981
SDR (Standard Deviation of Residuals):	0.107933
RAE (Relative Absolute Error):	0.99808
RRSE (Root Relative Squared Error [RMSE / Stdev(ref)]):	1.06387
MAPE (Mean Absolute Percentage Error):	0.262113

Elapsed Time: 00:00:49 Clear

nNetLite Simulator



THANK YOU

moshe.haiut@dspg.com

DSP®
GROUP



Premier Sponsor



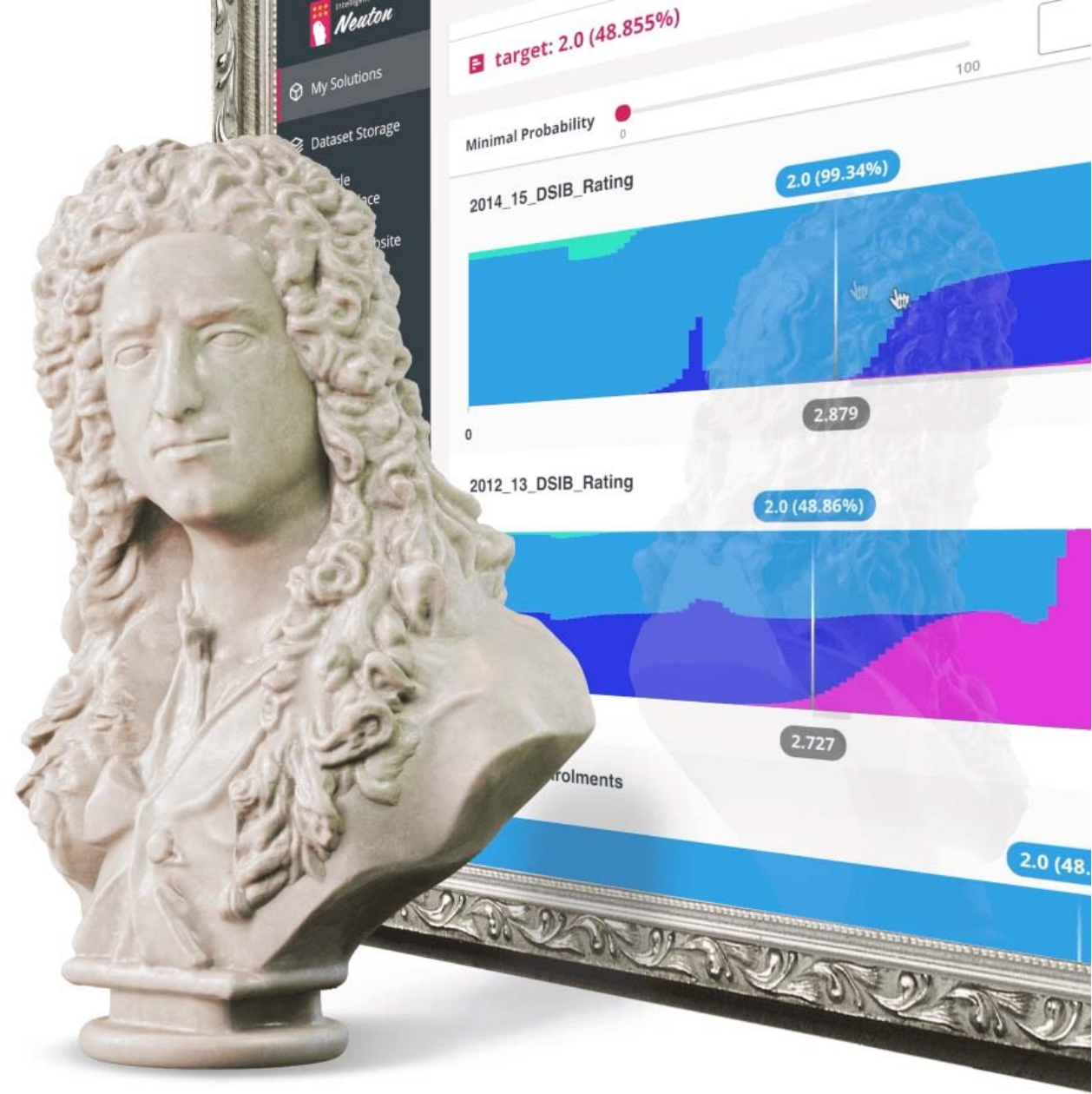
Automated TinyML

Zero-code SaaS solution

**Create tiny models, ready for embedding,
in just a few clicks!**

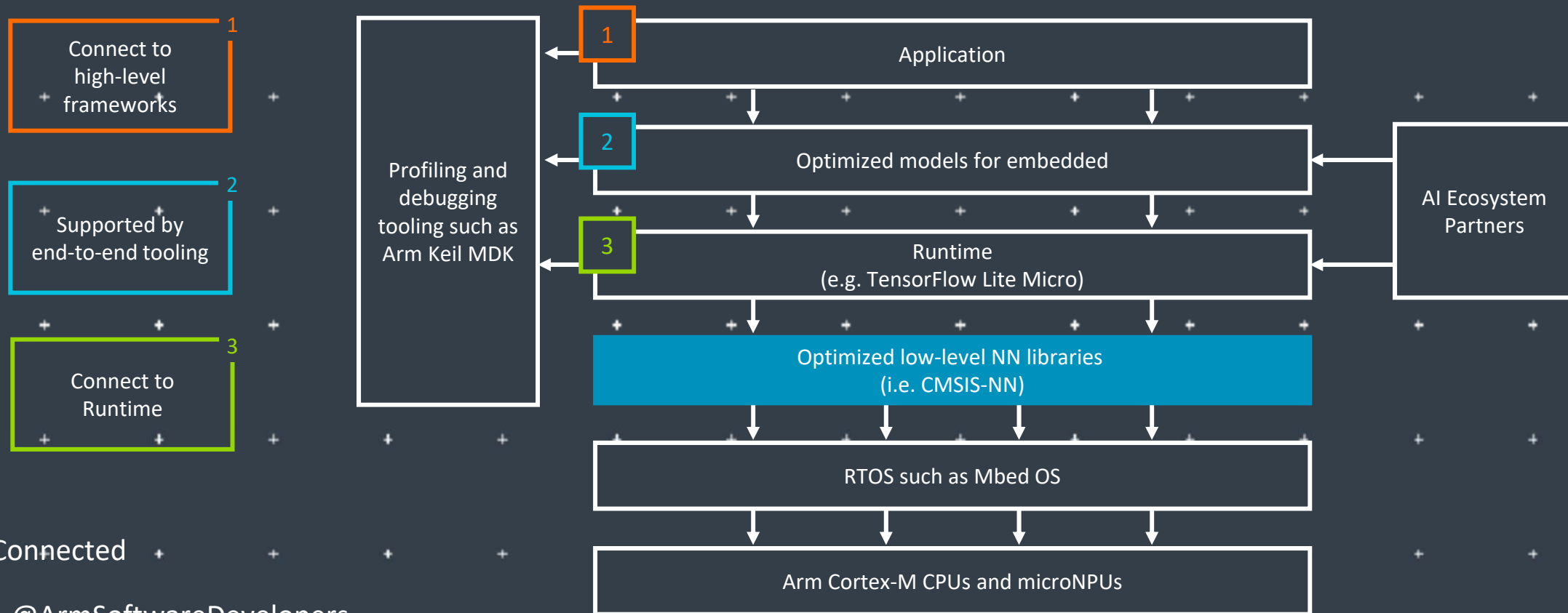
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.



Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



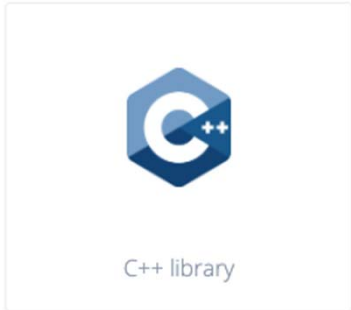
Stay Connected

 @ArmSoftwareDevelopers

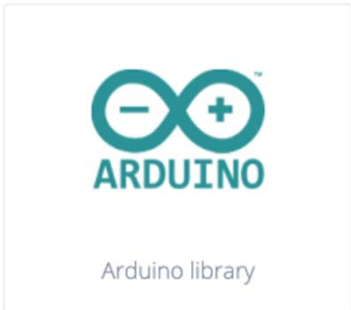
 @ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

TinyML for all developers



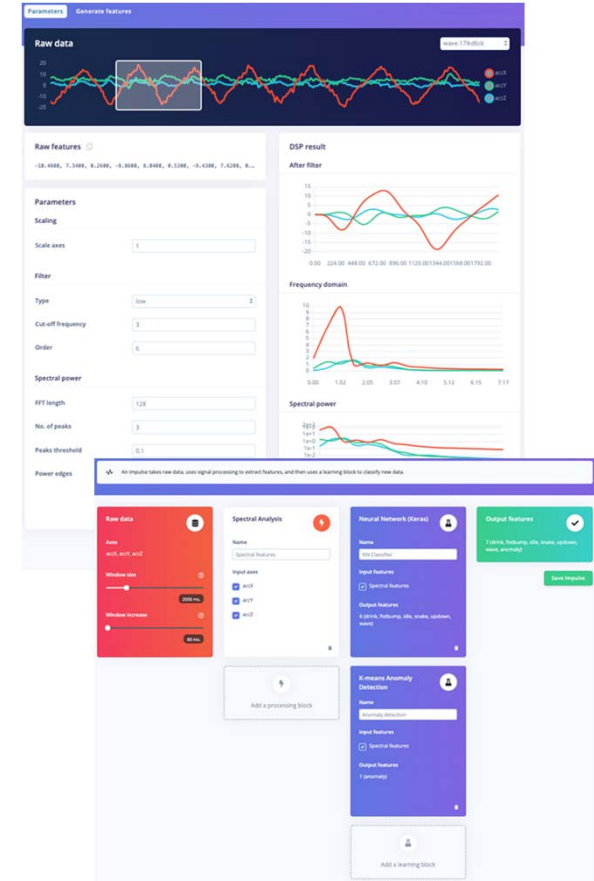
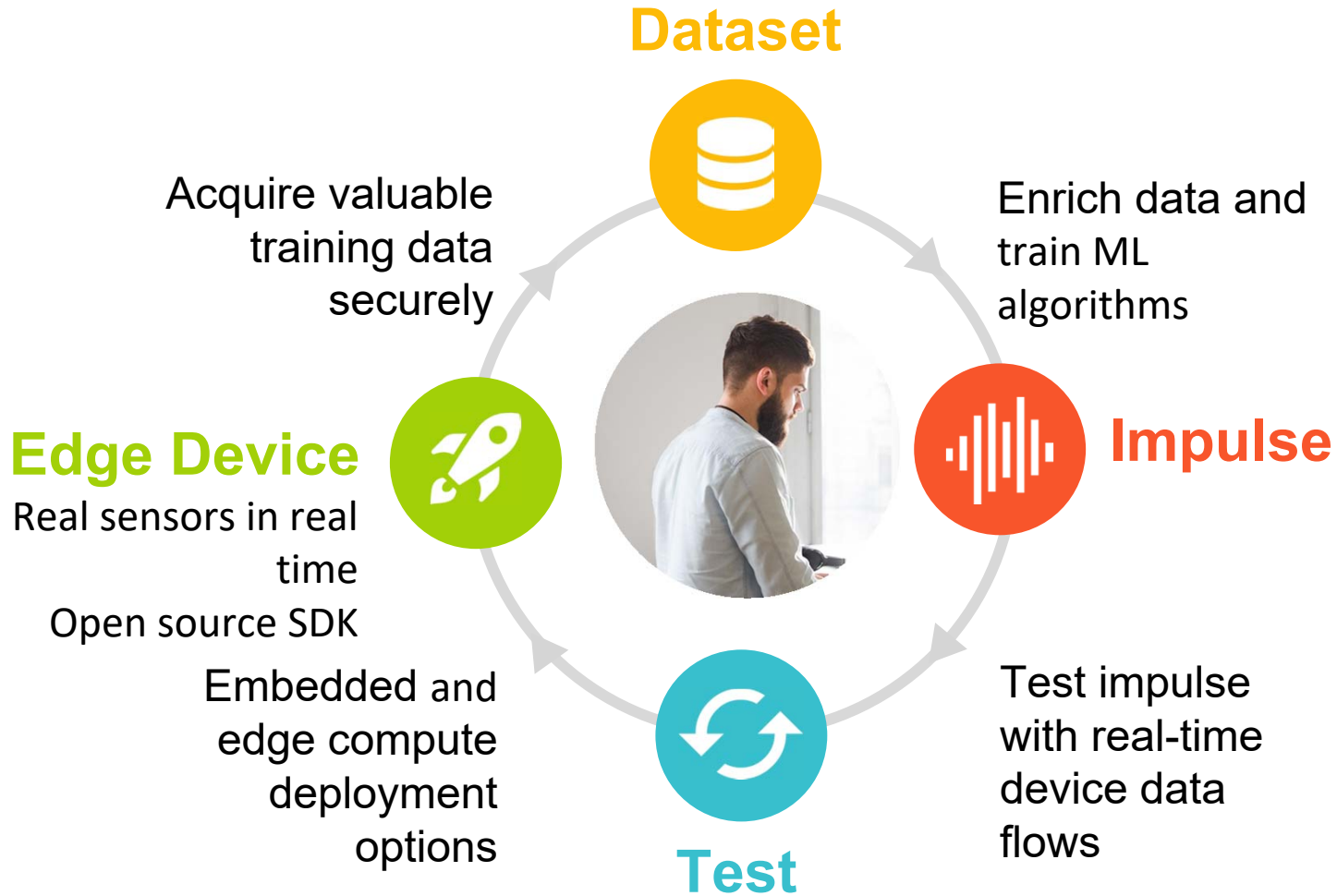
C++ library



Arduino library



WebAssembly



Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry



Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction



Action

Reinforcement learning for decision making



Edge cloud



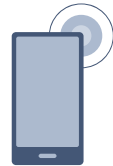
Cloud



IoT/IIoT



Automotive



Mobile

SYNTIANT

[Syntiant Corp.](#) is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a [CES® 2021 Best of Innovation Awards Honoree](#), [shipped over 10M units worldwide](#), and [unveiled the NDP120](#) part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com



@Syntiantcorp

Platinum Sponsors



Part of your life. Part of tomorrow.

www.infineon.com



Reality AI[®]

Add Advanced Sensing to your Product with Edge AI / TinyML

<https://reality.ai>



info@reality.ai



[@SensorAI](https://twitter.com/SensorAI)



[Reality AI](https://www.linkedin.com/company/reality-ai)

**Pre-built Edge AI sensing modules,
plus tools to build your own**

Reality AI solutions

Prebuilt sound recognition models for
indoor and outdoor use cases

Solution for industrial anomaly detection

Pre-built automotive solution that lets cars
“see with sound”

Reality AI Tools[®] software

Build prototypes, then turn them into
real products

Explain ML models and relate the function
to the physics

Optimize the hardware, including
sensor selection and placement



Gold Sponsors



Latent AI

Adaptive AI for the Intelligent Edge

[Latentai.com](https://latent.ai)



Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

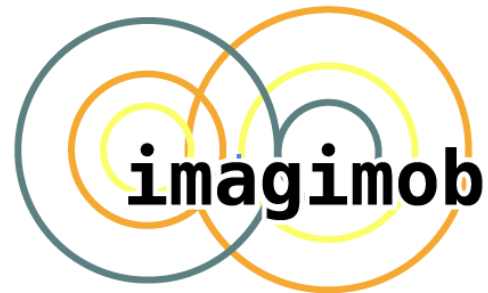
- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.



sensiml.com

Silver Sponsors





Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® EMEA Technical Forum 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at tinyML EMEA. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org