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

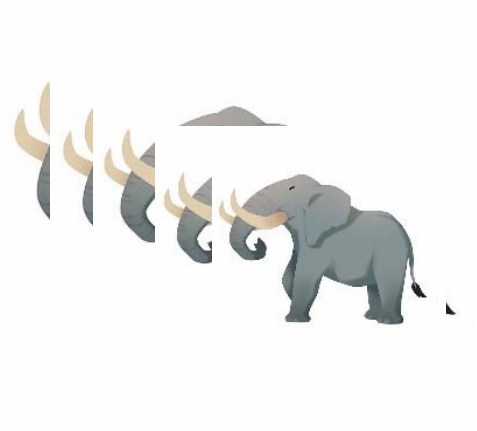
DSP Group
Innovative Minimization of Parameter Memory in small silicon low power devices

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

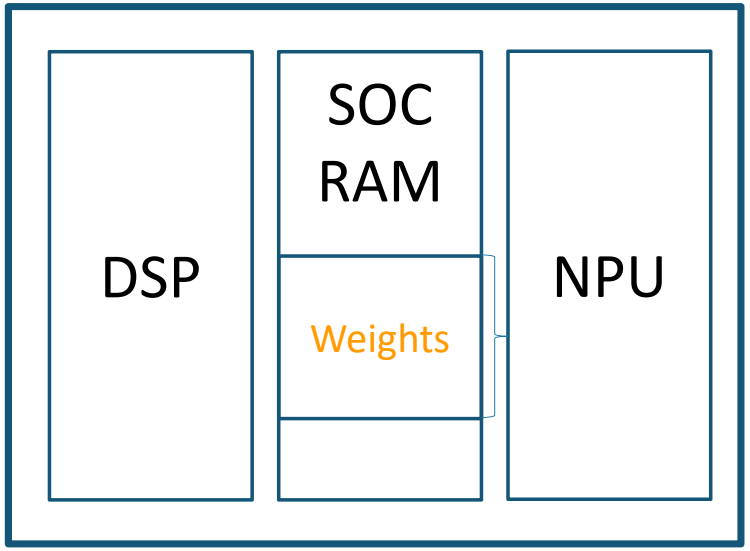


Quantization

Pruning

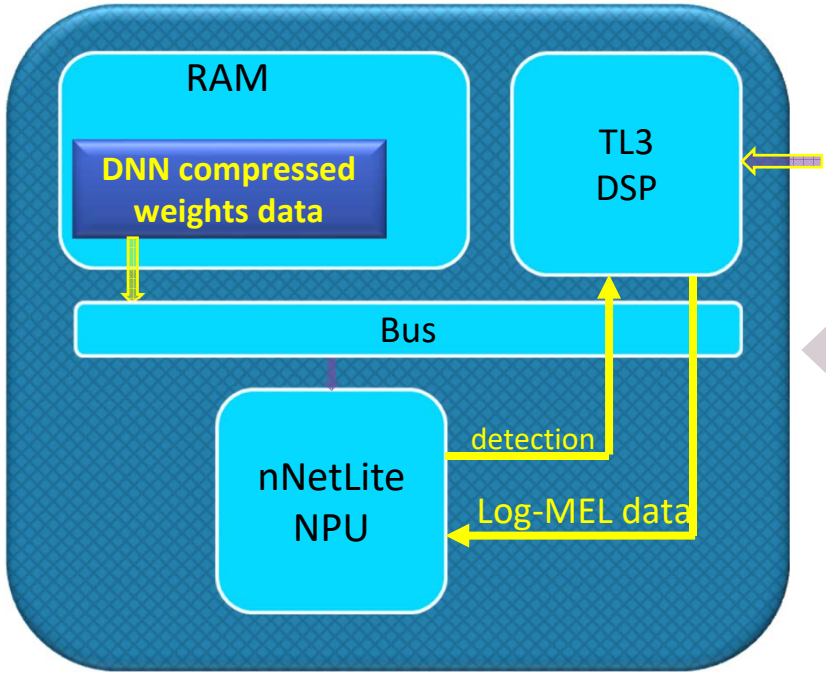
Entropy Coding

Packing



The DSPG nNetLite h/w Engine

- The nNetLite engine is a stand-alone module, planed 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



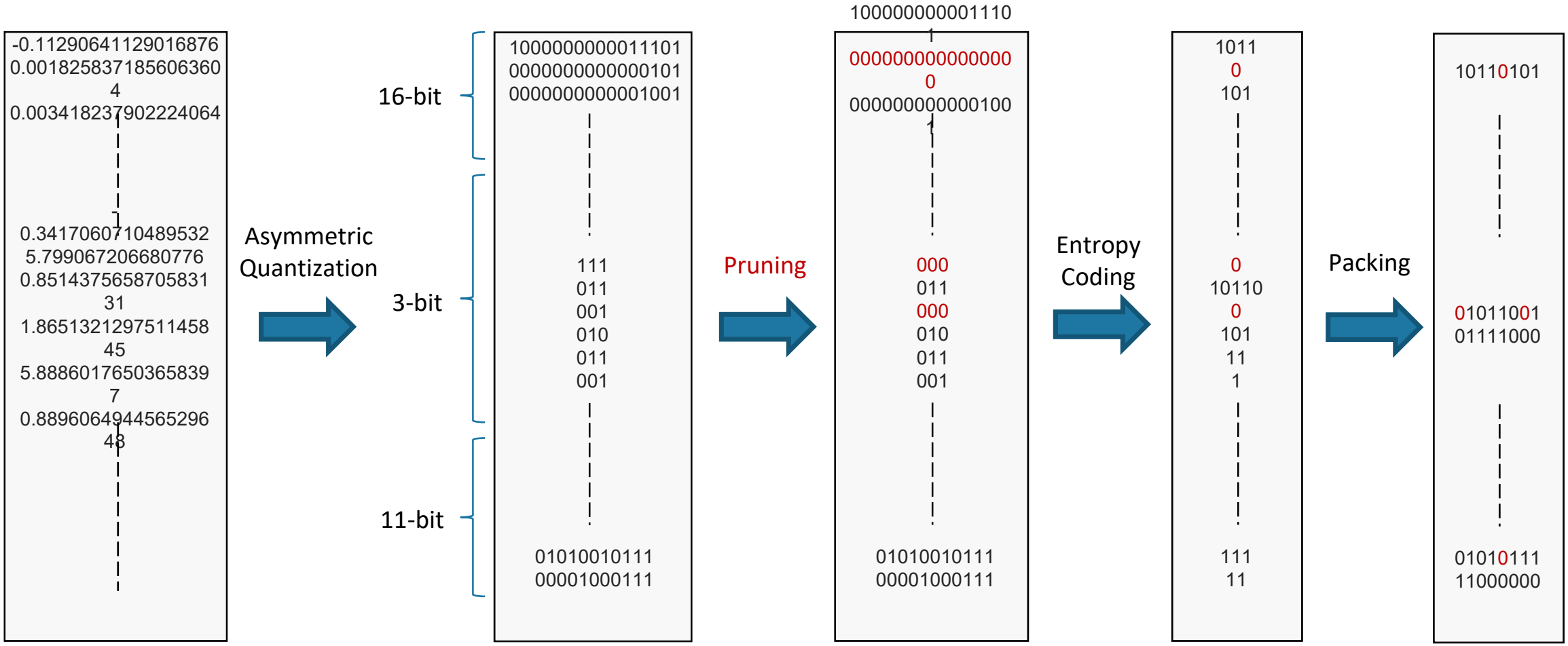
DBM10L incorporates the nNet Lite engine and a DSP

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

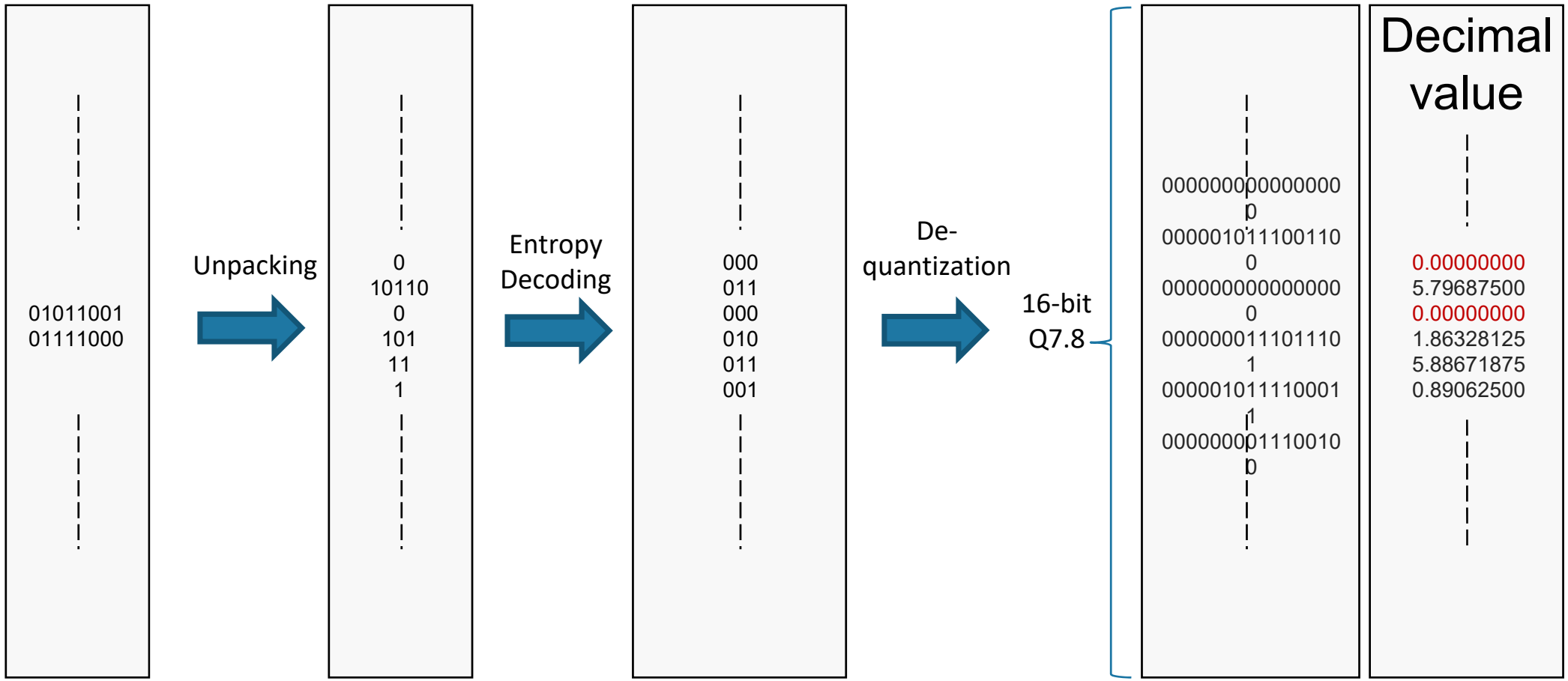
The DSP makes the audio pre-processing (e.g. log-MEL calc.) and the detection post processing to support the concurrent inferences

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

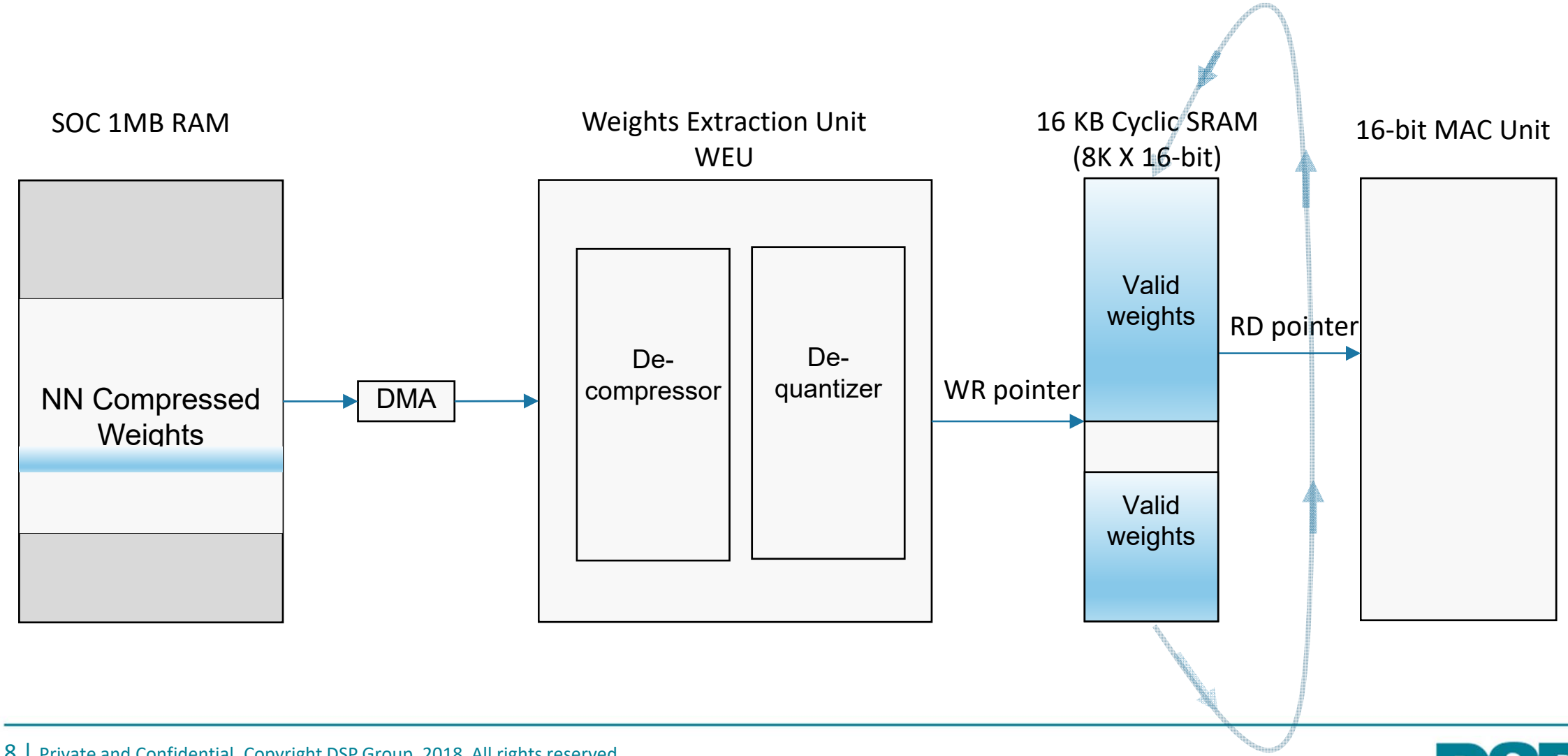
Weights Compression by the nNetLite Compiler



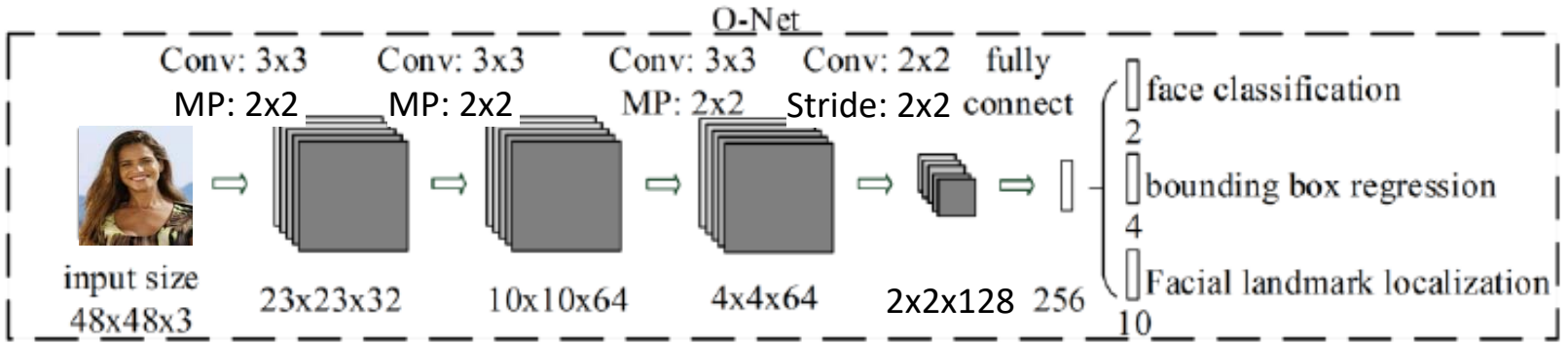
Weights De-compression by the nNetLite WEU h/w module

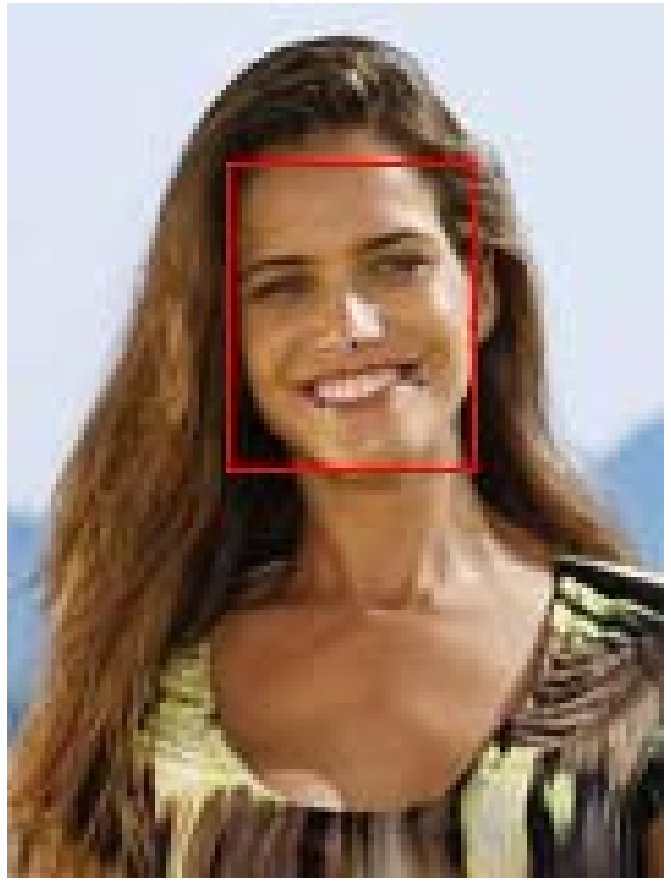


Weights De-compression by the nNetLite WEU h/w module



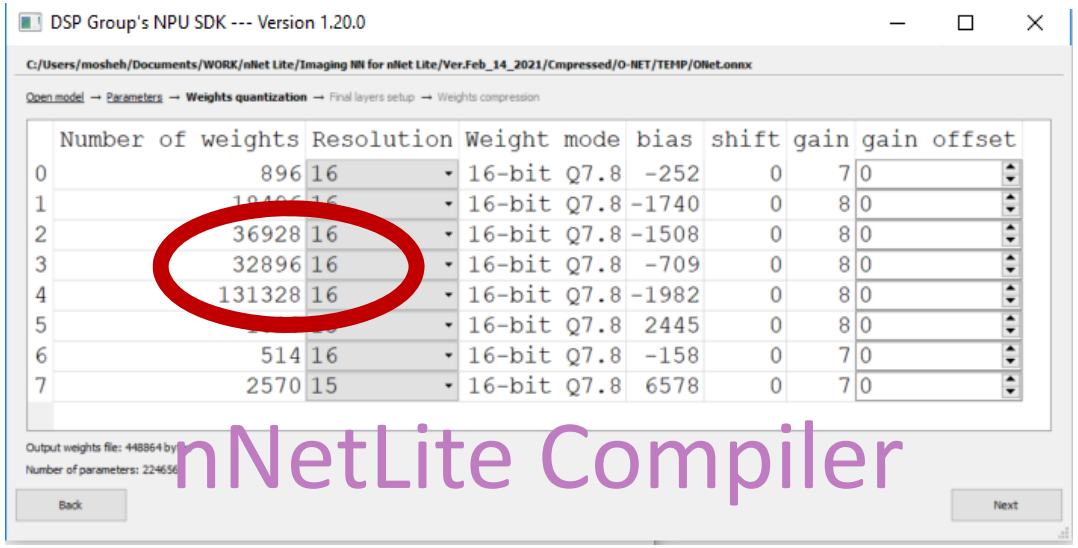
Example: Face Detection NN



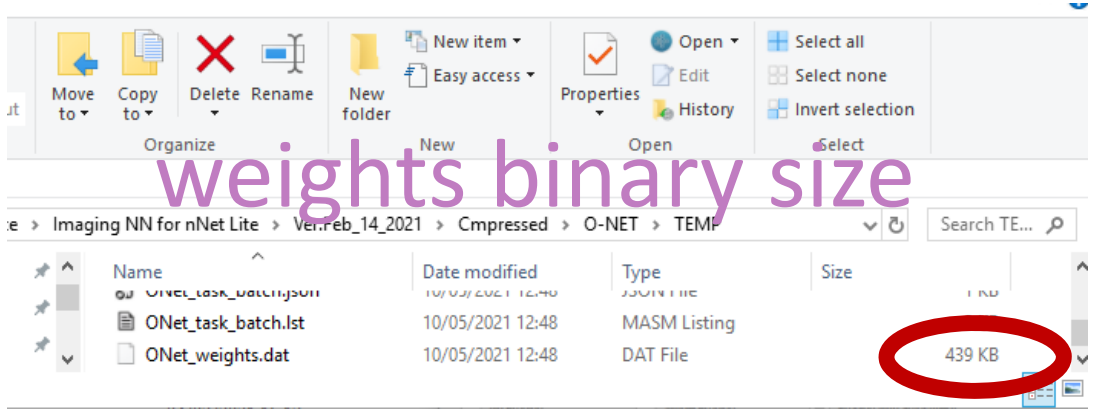


- Number of layers: 8
- Number of MAC ops: ~12M / Inference
- Number of Weights: 224,656
- Available memory: 280KB

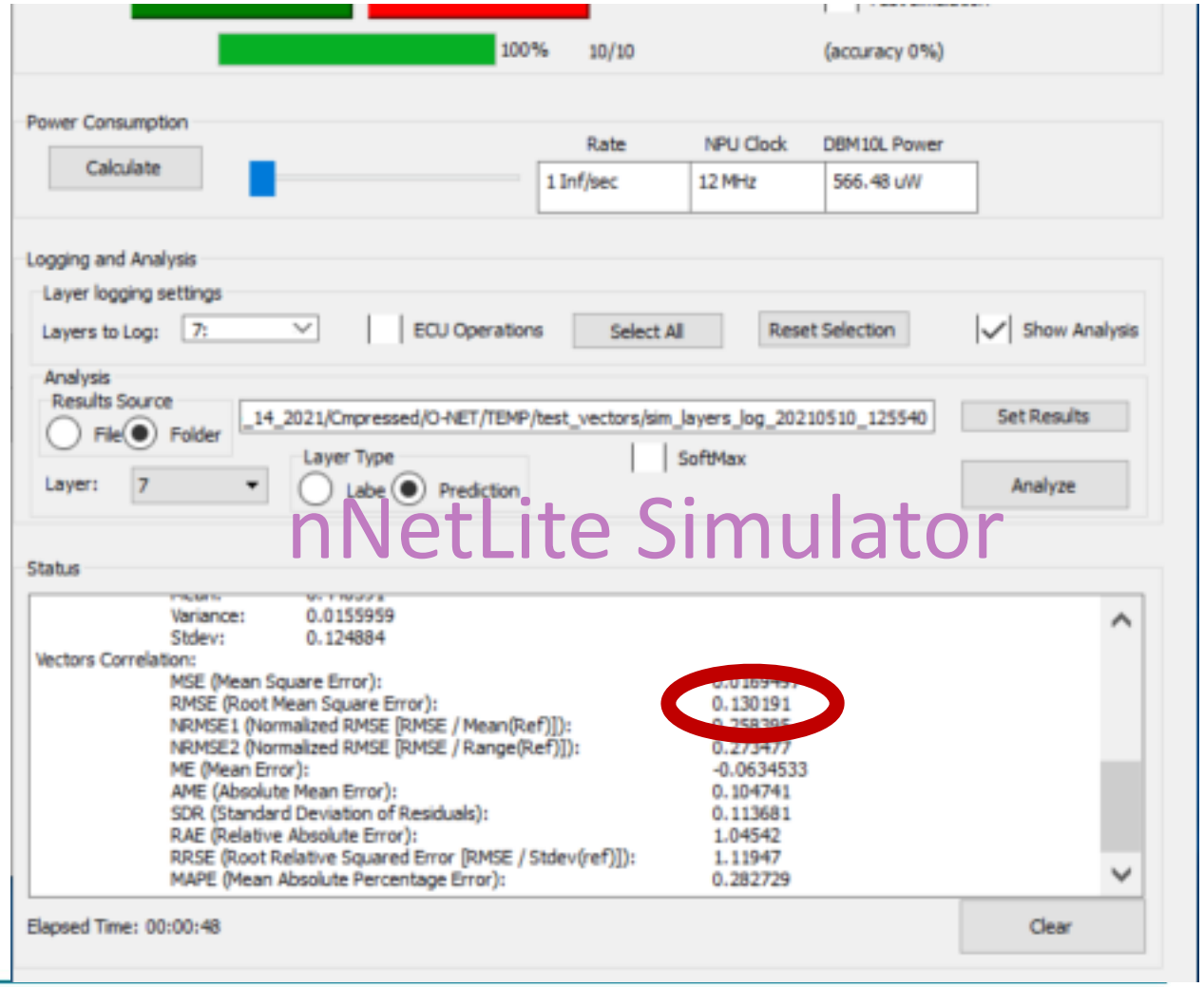
w/o Compression. Weights: 439 KB, RMSE: 0.130191



nNetLite Compiler

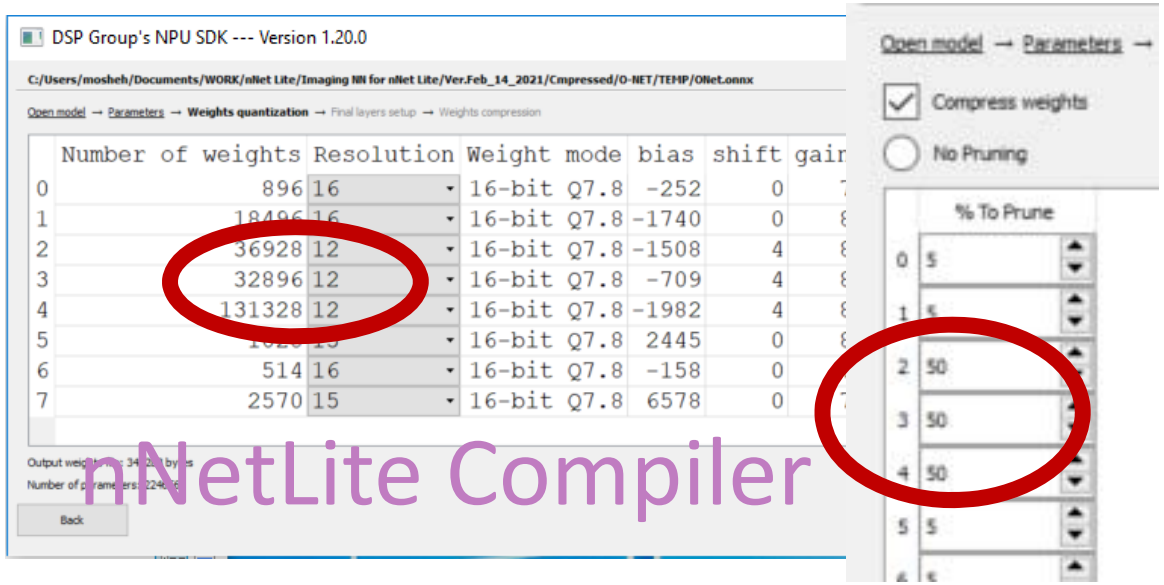


weights binary size

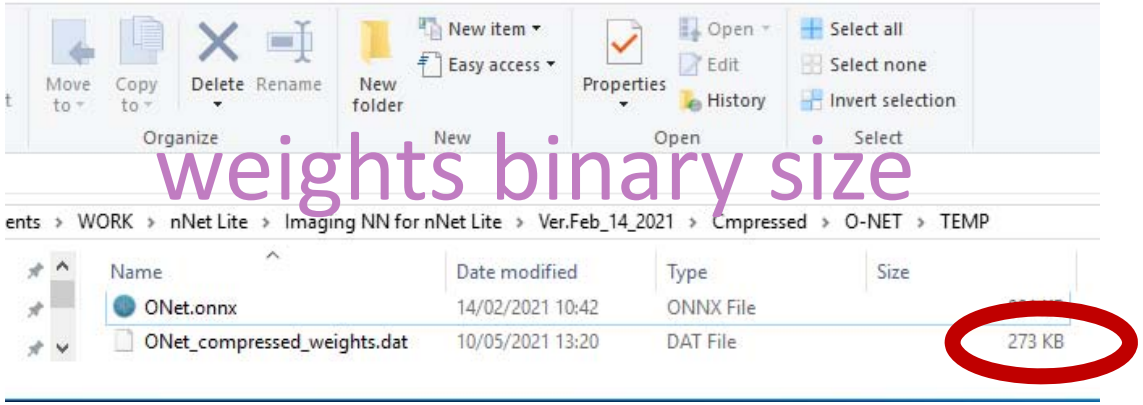


nNetLite Simulator

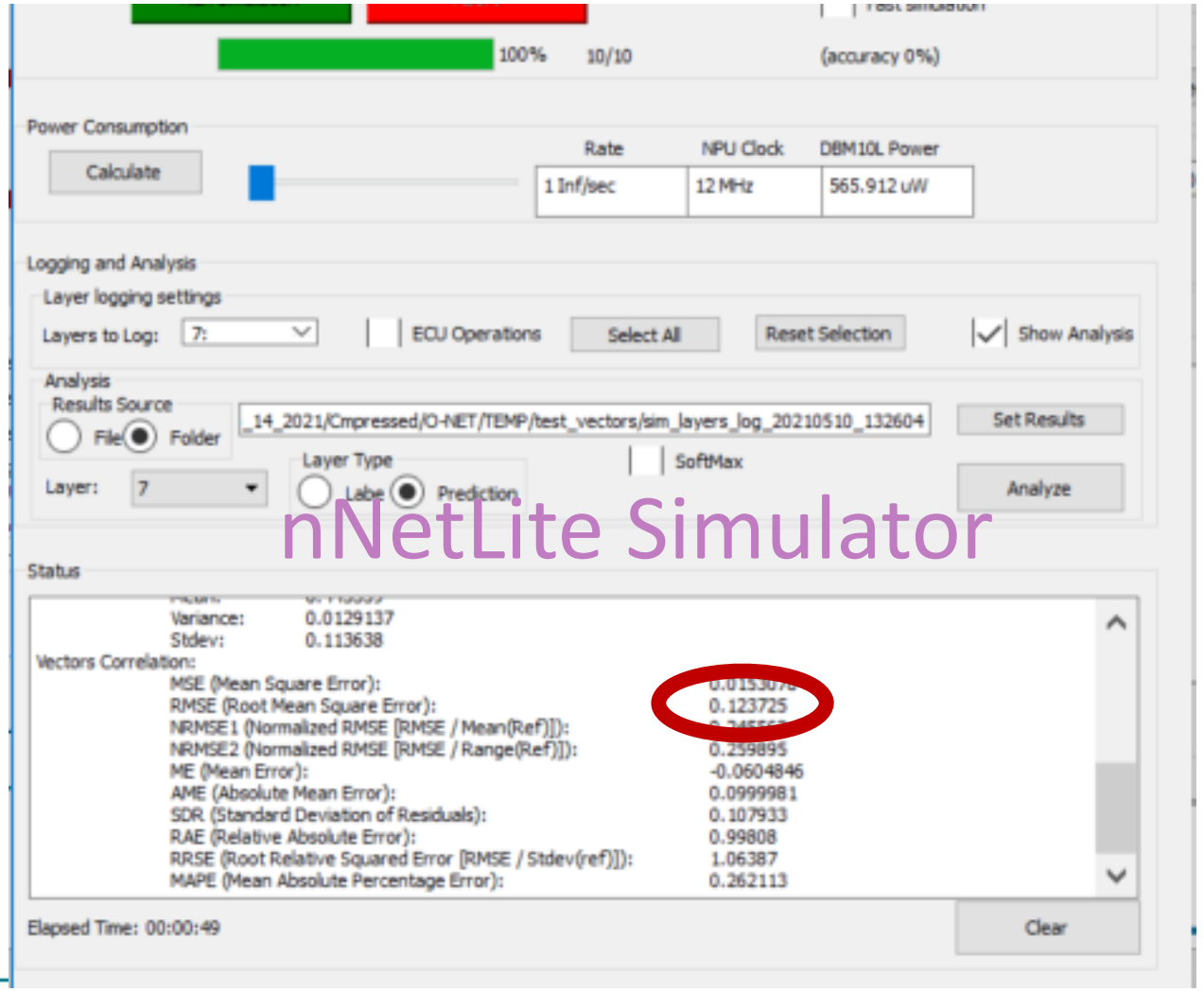
12-bit weights, 50% Prun. Weights: 273 KB, RMSE: 0.123725



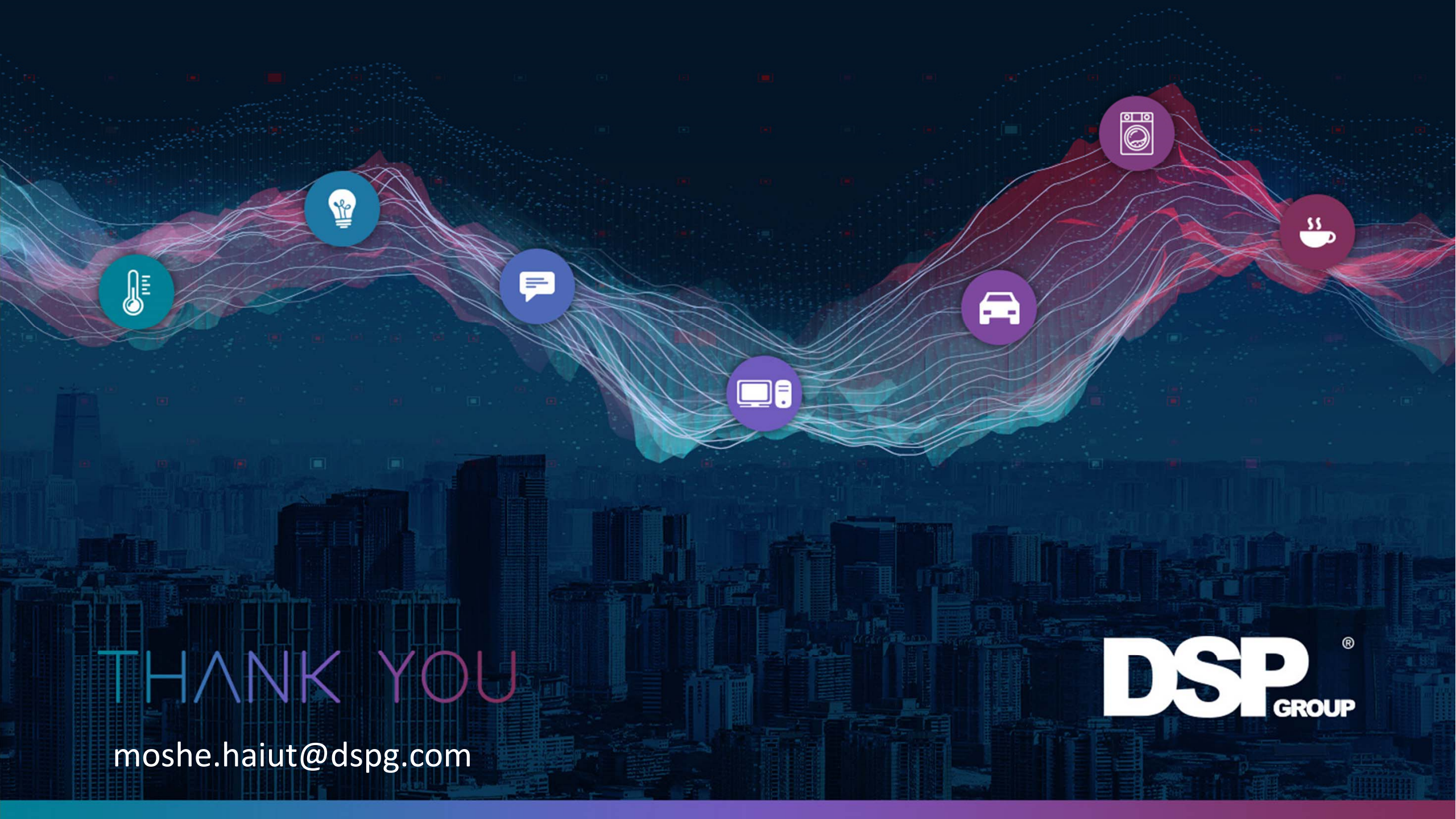
nNetLite Compiler



weights binary size



nNetLite Simulator



moshe.haiut@dspg.com

DSP GROUP®

TINY ML EMEA
Virtual
JUNE 7-10,2021

Premier Sponsor

Intelligent Agent
Neuton

Automated TinyML

Zero-code SaaS solution

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

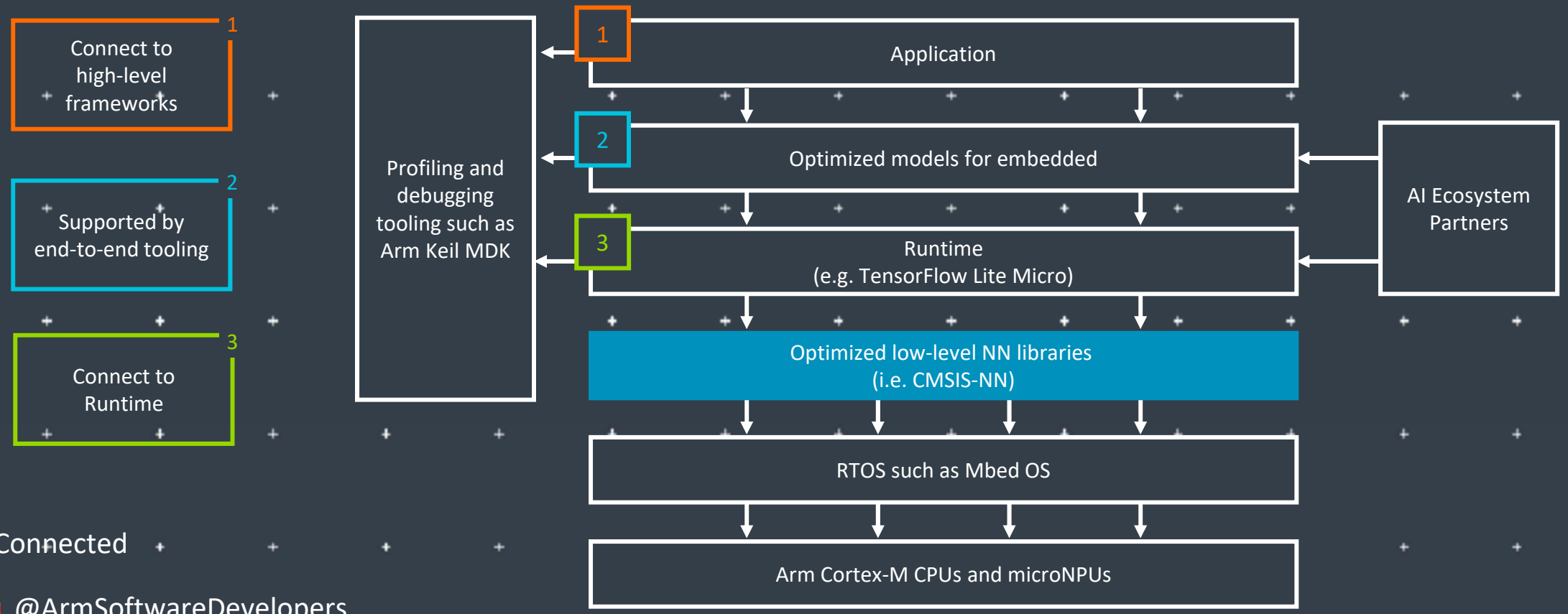
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

TINY ML EMEA
Virtual
JUNE 7-10,2021

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

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

arm

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

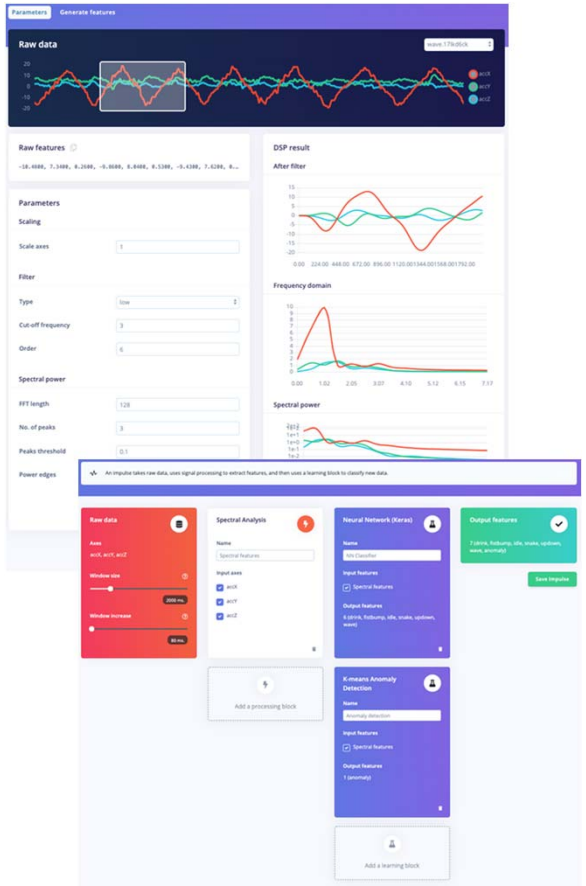
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

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

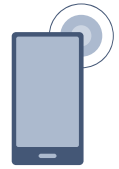
Reinforcement learning for decision making

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

TINY ML EMEA
Virtual
JUNE 7-10,2021

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 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

TINY ML EMEA
Virtual
JUNE 7-10,2021

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

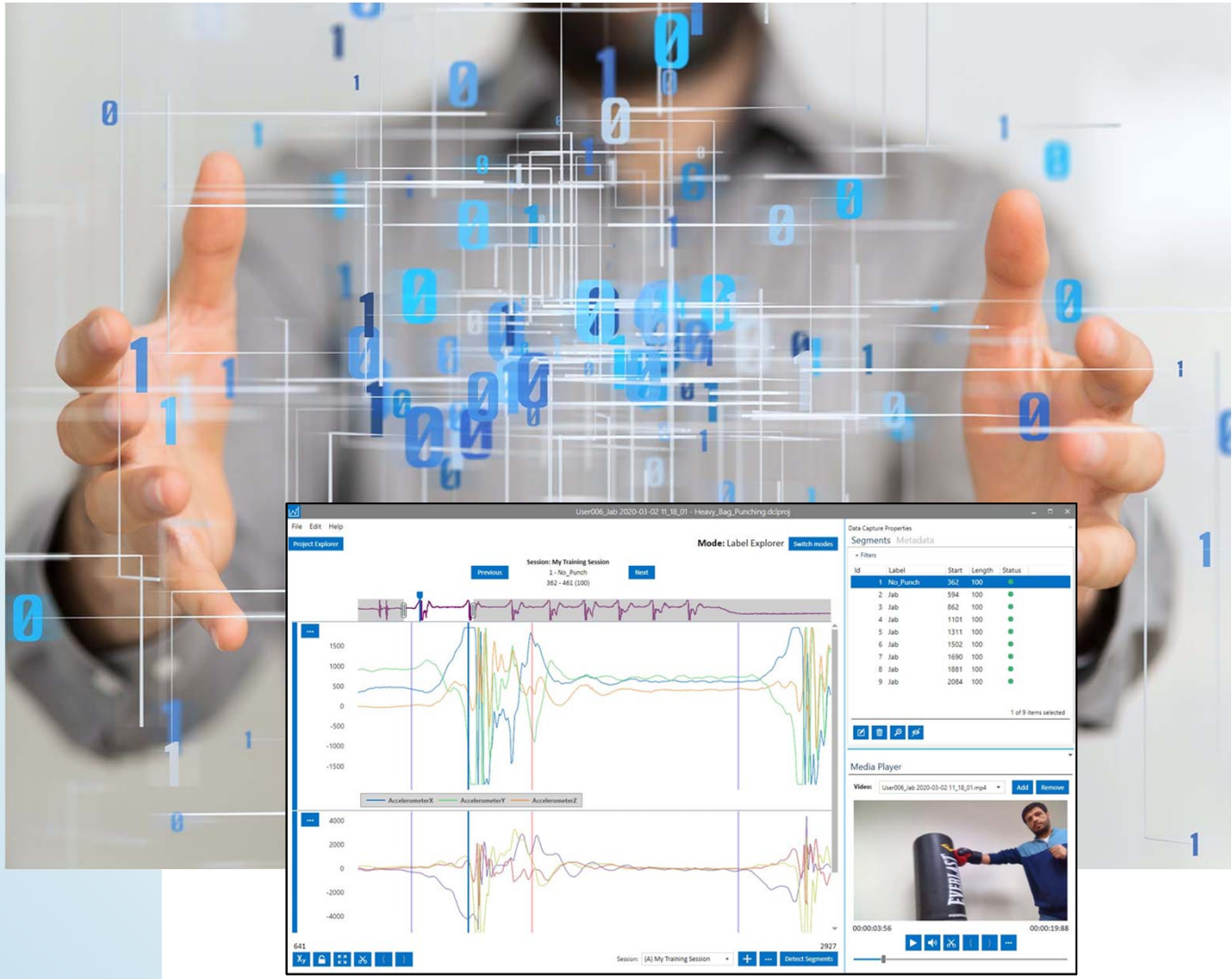
SensiML™

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

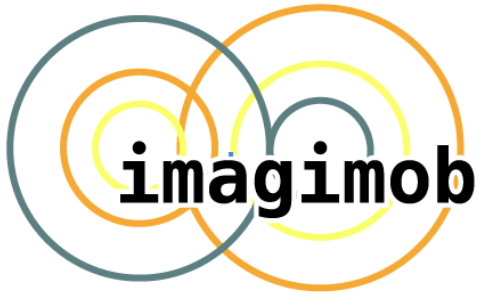
We enable the creation of production-grade smart sensor devices.



sensiml.com

TINY ML EMEA Virtual JUNE 7-10,2021

Silver Sponsors



Copyright Notice