# tinyML® Talks

*Enabling Ultra-low Power Machine Learning at the Edge*

## "Universal CNN Accelerator Intended for Edge-Based AI Inference"

Rastislav Struharik - University of Novi Sad

Germany Area Group – April 7, 2021

www.tinyML.org

# tinyML Talks Sponsors

*tinyML Strategic Partner*

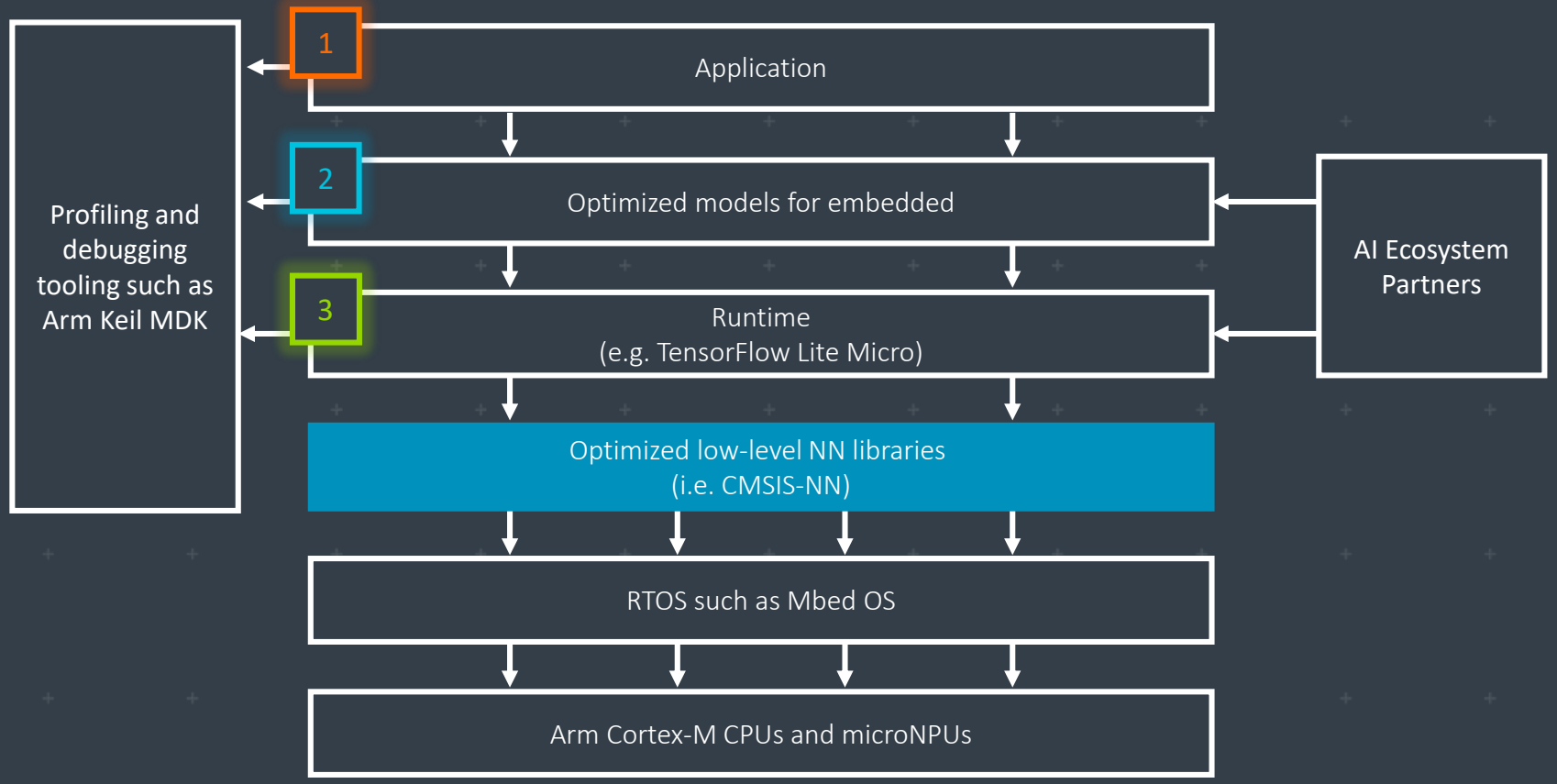
Additional Sponsorships available – contact Olga@tinyML.org for info

# Arm: The Software and Hardware Foundation for tinyML

1 Connect to high-level frameworks

2 Supported by end-to-end tooling

3 Connect to Runtime




Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

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




arm

Deeplite

# WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

**Automatically compress** SOTA models like MobileNet to <200KB with **little to no drop in accuracy** for inference on resource-limited MCUs

**Reduce** model optimization trial & error from weeks to days using Deeplite's **design space exploration**


**Deploy more** models to your device without sacrificing performance or battery life with our **easy-to-use software**

BECOME BETA USER **bit.ly/testdeeplite**

mobilityXlab arm AI 100 CBINSIGHTS

# TinyML for all developers

C++ library

Arduino library

WebAssembly

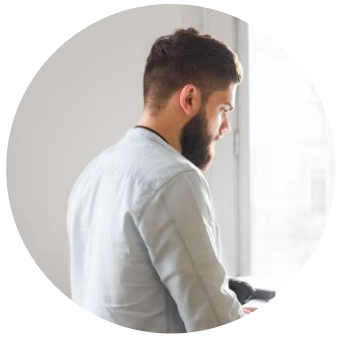
## Dataset

Acquire valuable training data securely

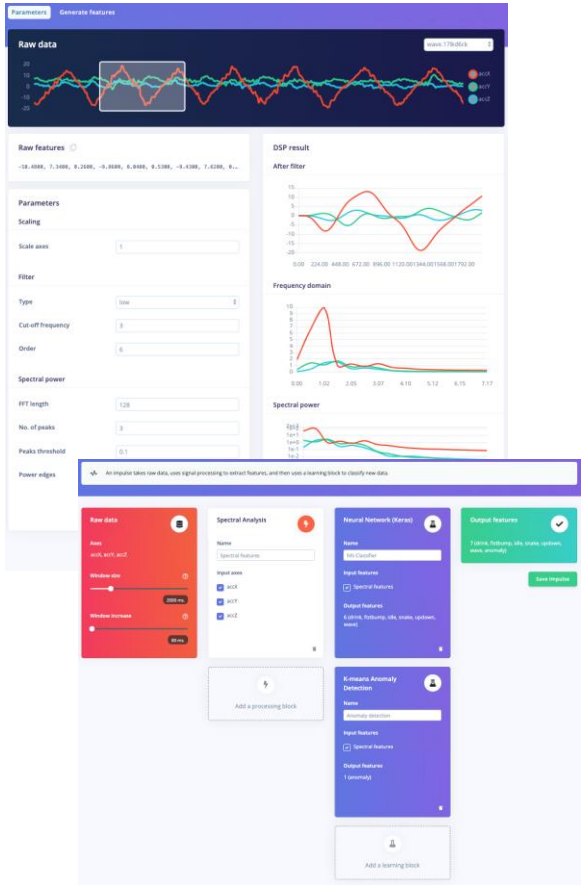
Enrich data and train ML algorithms

## 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

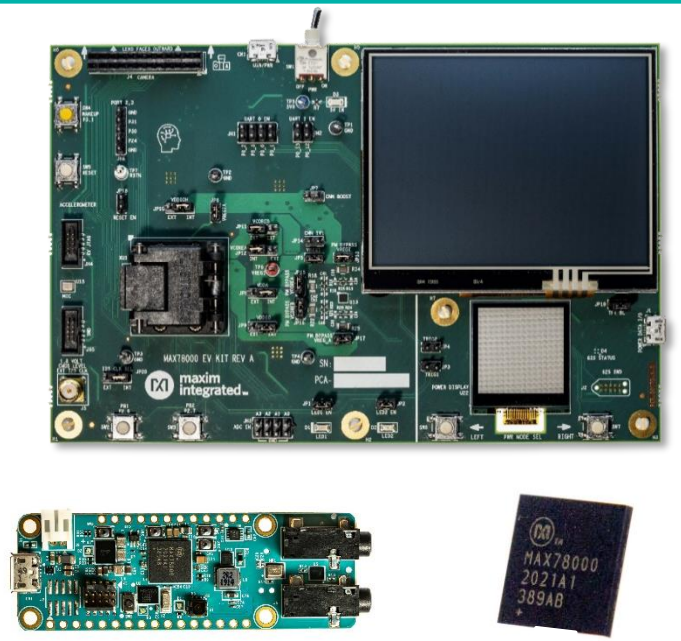
## Test



www.edgeimpulse.com

# Maxim Integrated: Enabling Edge Intelligence

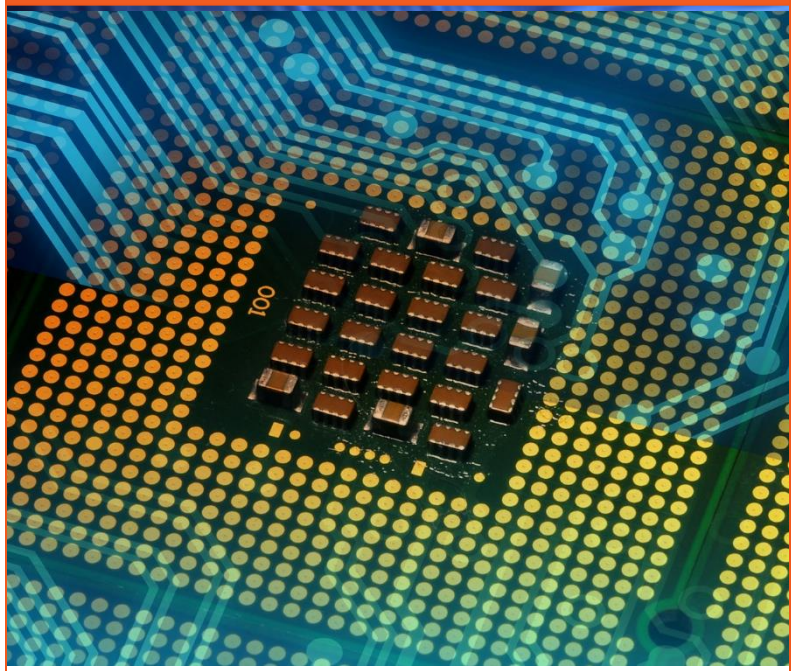
## Advanced AI Acceleration IC




The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000

## Low Power Cortex M4 Micros



Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

## Sensors and Signal Conditioning


Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors

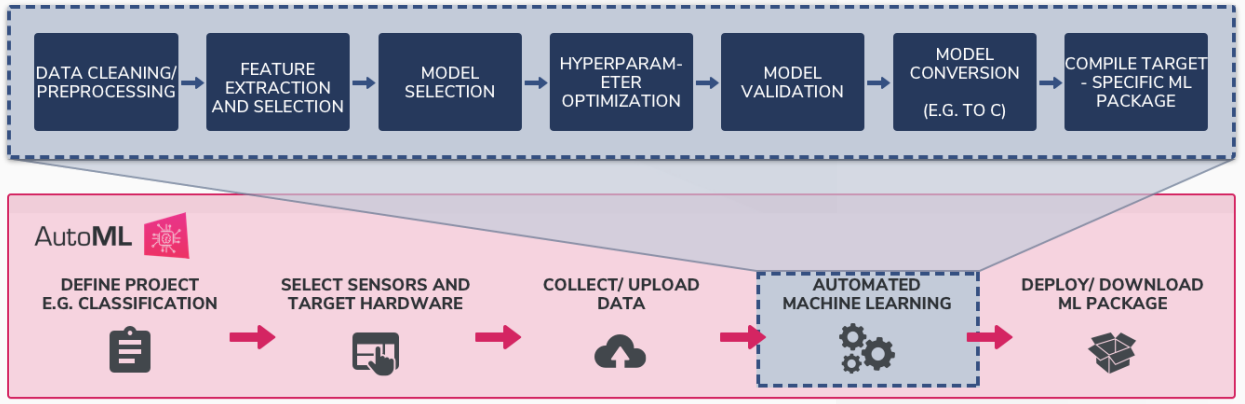
# Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

## Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

## End-to-End Machine Learning Platform




**For more information, visit: www.qeexo.com**

## Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

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

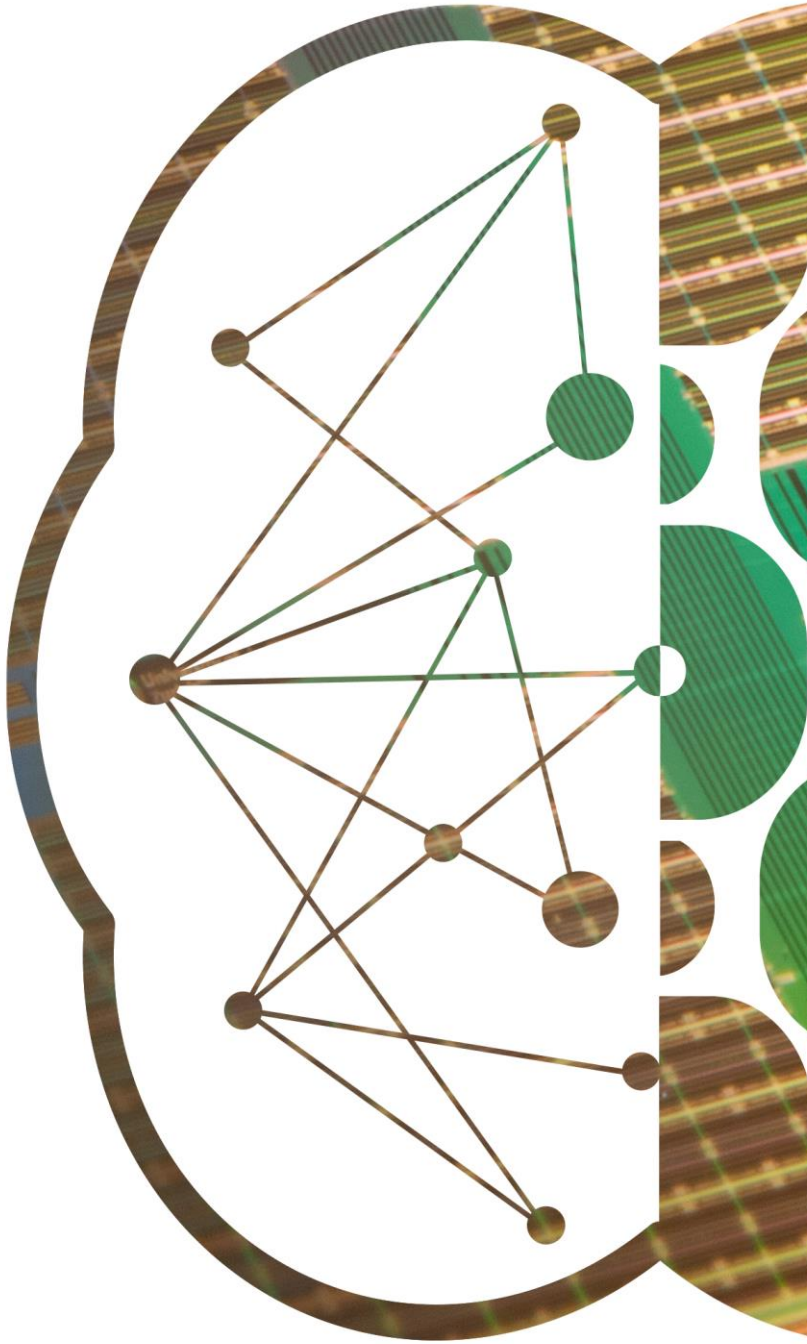
Explain ML models and relate the function to the physics

Optimize the hardware, including sensor selection and placement

# SynSense

**SynSense** builds **sensing and inference** hardware for **ultra-low-power** (sub-mW) **embedded, mobile and edge** devices. We design systems for **real-time always-on smart sensing**, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai



TINY ML TALKS webcast

# KIM-Labs

Our offer for small and medium-sized enterprises (SMEs):

- AI trainings
- AI events
- AI development projects for your application
- AI Working Groups & Expert Tables
- AI Co-Working Spaces & Prototyping Trainings

KIM-Labs

MedicalMountains

HOCHSCHULE FURTWANGEN UNIVERSITY HFU

TECHNOLOGYMOUNTAINS
Der Technologieverbund im Südwesten

Hahn Schickard
Visions to Products

IW4.0 Initiative Wirtschaft 4.0 BW

digital@bw

Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

# Next tinyML Talks

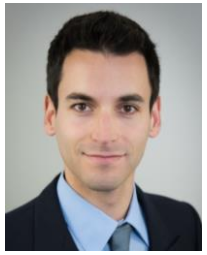
| Date | Presenter | Topic / Title |
|---|---|---|
| Tuesday,<br>April 13 | **Bernhard Suhm**<br>Machine Learning Product Manager,<br>MathWorks | Deploying AI to Embedded Systems |

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting

# Local Committee in Germany

Alexis Veynachter,
Master Degree in Control Engineering, Senior Field Application Engineer
Infineon 32bits MCUs for Sensors, Fusion & Control



Carlos Hernandez-Vaquero
Software Project Manager, IoT devices
Robert Bosch

Prof. Dr. Daniel Mueller-Gritschneder
Interim Head - Chair of Real-time Computer Systems
Group Leader ESL - Chair of Electronic Design Automation
Technical University of Munich

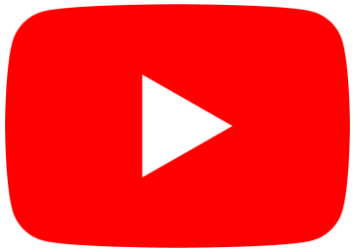
Marcus Rüb
Researcher in the field of TinyML
Hahn-Schickard

TINY ML TALKS webcast

# Reminders

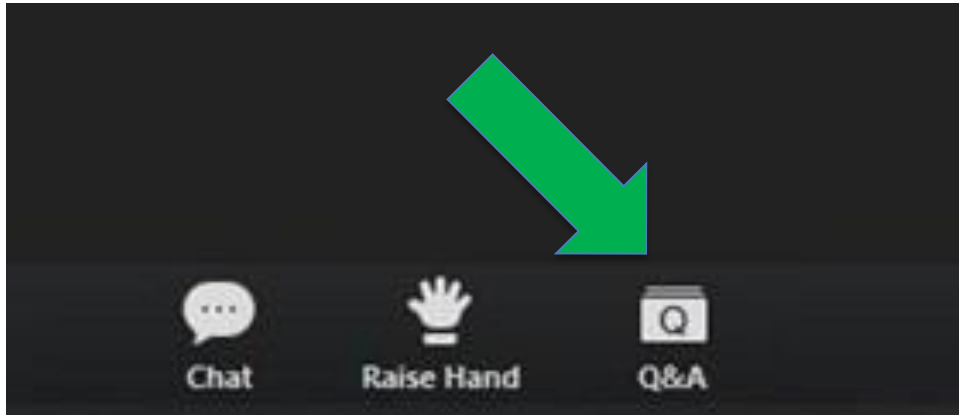
Slides & Videos will be posted tomorrow


tinyml.org/forums



youtube.com/tinyml

Please use the Q&A window for your questions



TINY ML TALKS webcast

# Rastislav Struharik

Rastislav Struharik is a full professor at the Department of Power, Electronics and Telecommunications, Faculty of Technical Sciences, University of Novi Sad, Serbia. He received his PhD in Electronics in 2009, in the area of hardware acceleration of machine learning algorithms. During his academic career he has published more than 35 papers in international journals and conferences, mainly focusing on the hardware acceleration of machine learning algorithms, such as Decision Trees, Support Vector Machines, Artificial Neural Networks, Convolutional Neural Networks, and Ensemble Classifiers, targeting both learning and inference algorithms. For the past three years he has also been working as the chief architect for the IDS own FPGA IP core technology for hardware acceleration of Convolutional Neural Networks, intended for edge AI applications.

# IT'S SO EASY.

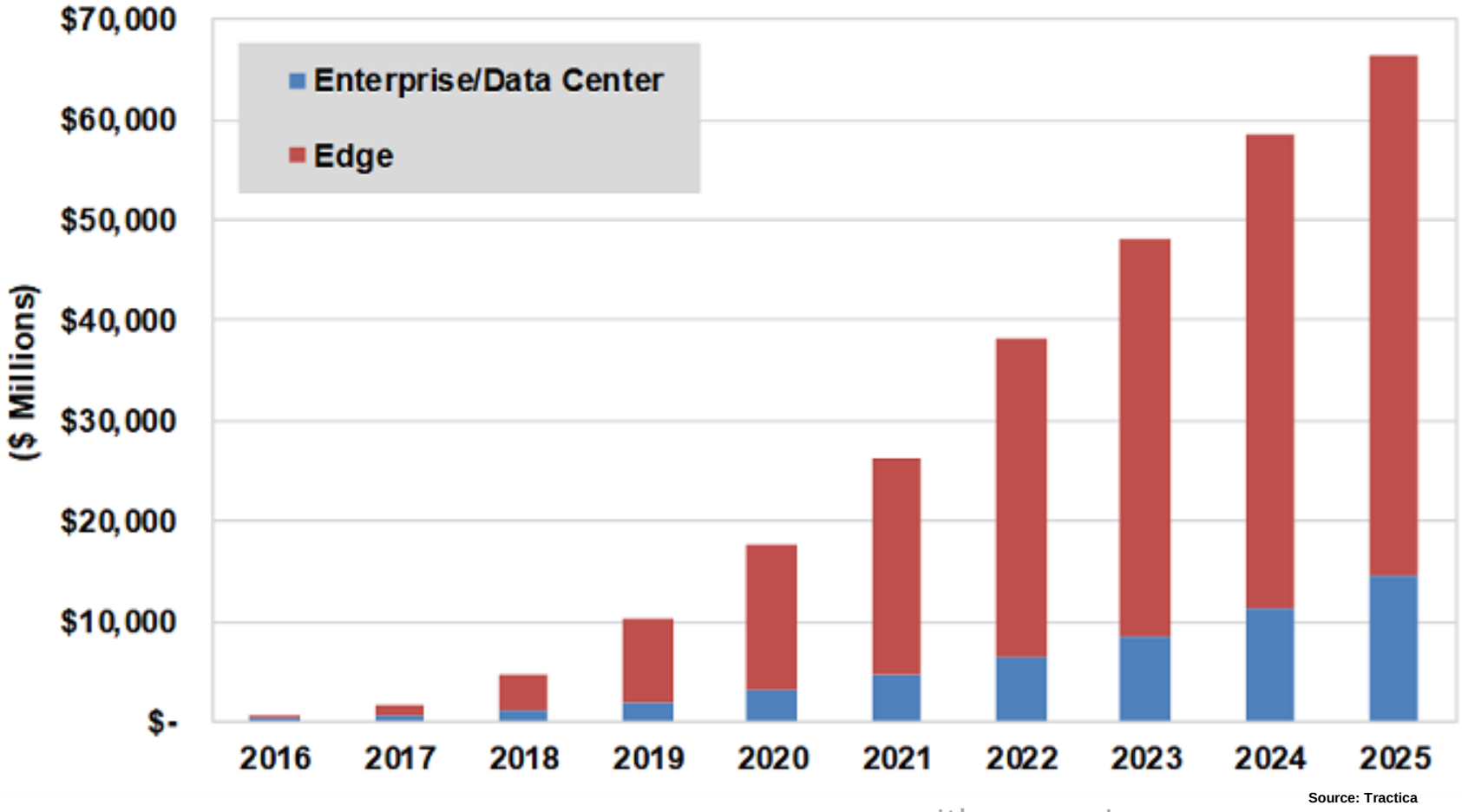
WITH IDS IMAGING SOLUTIONS



## Universal CNN Accelerator for Edge-based AI Inference

Rastislav Struharik

# Edge AI – The next big thing?




**AI processing**

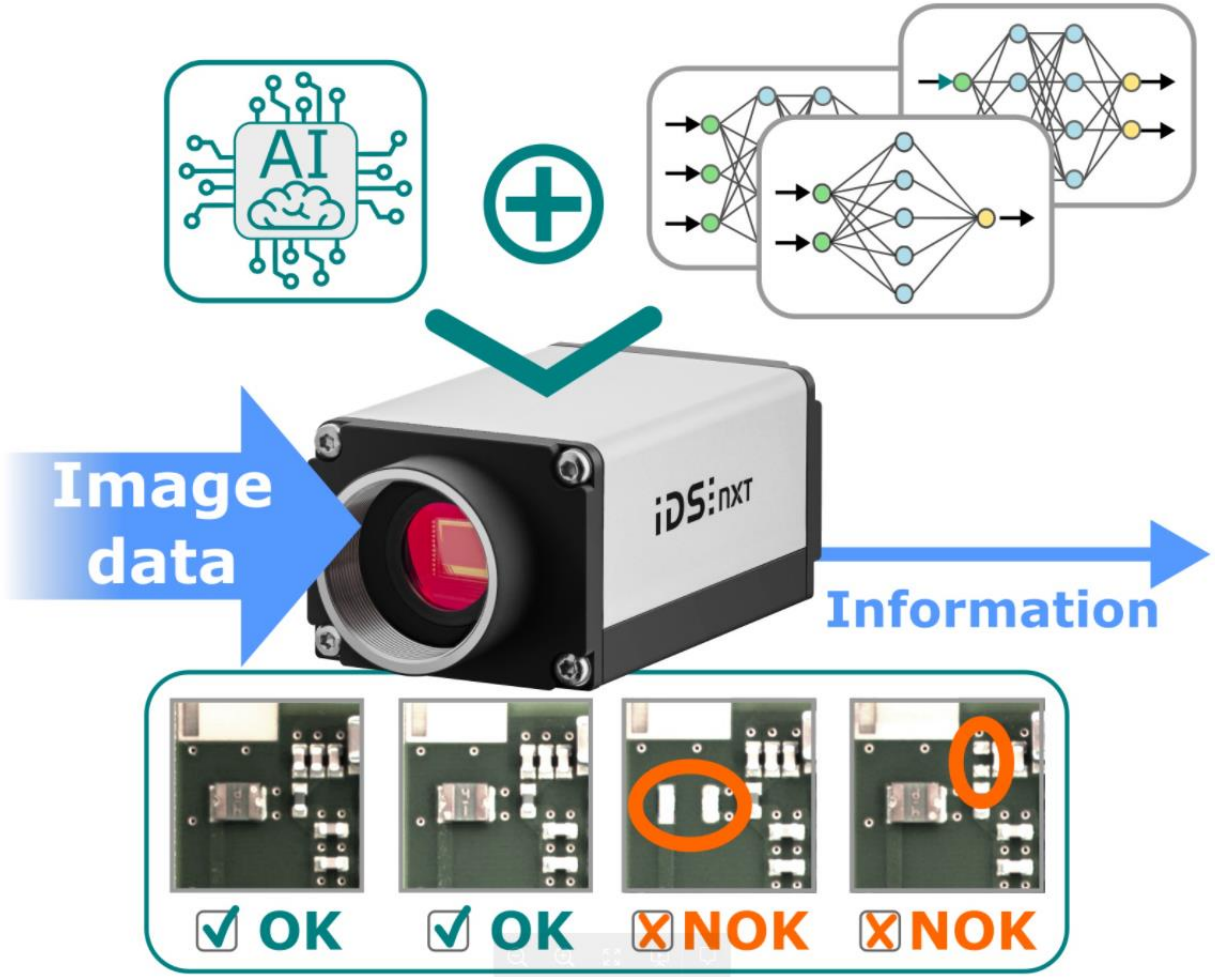
→ on the Edge, where things are going to get interesting

**AI training**

→ will stay in the cloud, where Nvidia's GPUs are most suited for the job

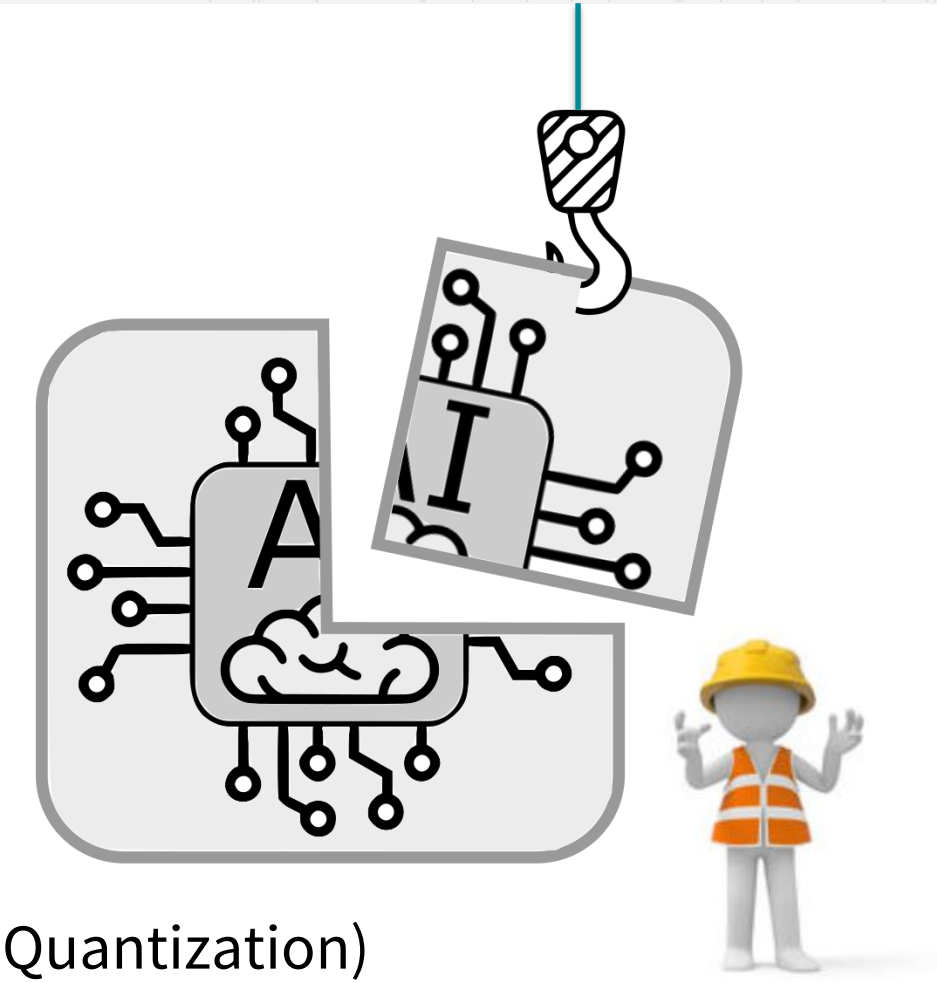
# Why AI on the edge?

- Bandwidth and Latency
- Security and Privacy
- Reliability and Availability
- Customization

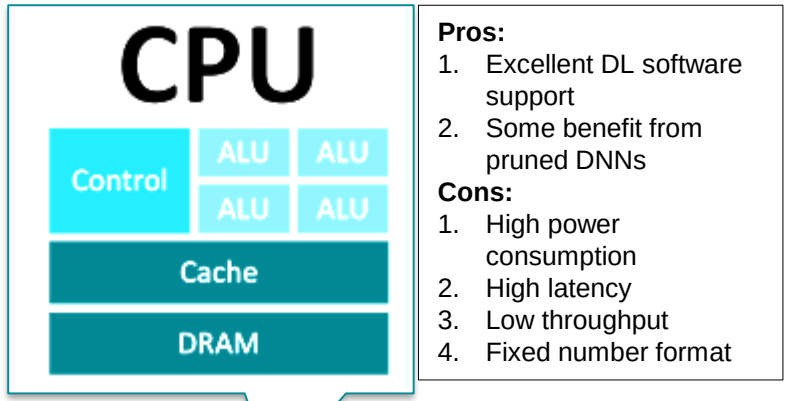


# Challenges in deploying AI on the edge

- **Edge Computing**
  - is all about efficiency
  - requires low latency
- **Edge AI requires:**
  - Parameter Efficient Networks
    (MobileNet, EfficientNet, SqueezeNet, …)
  - Network Compression (by using Pruning and Quantization)
  - Customized Computing Systems



# Available options when implementing CNNs on the edge




**Pros:**
1. Excellent DL software support
2. Some benefit from pruned DNNs

**Cons:**
1. High power consumption
2. High latency
3. Low throughput
4. Fixed number format

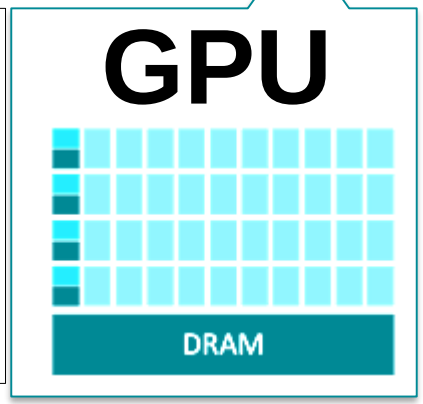

**Pros:**
1. Extremely power efficient
2. Low latency and high throughput
3. Some support of different number formats

**Cons:**
1. High developemnt cost and time
2. Limited DL software support
3. No support for pruned DNNs
4. Diffcult to support new DL architectures

**Flexibility** ⟷ **Efficiency**

**Pros:**
1. Excellent DL software support
2. High throughput
3. Some support of different number formats

**Cons:**
1. High power consumption
2. High latency
3. Fixed number format
4. Doesn‘t support pruned DNNs
5. Short lifespan



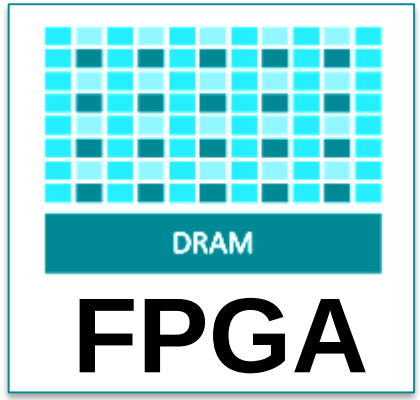


**Pros:**
1. Power efficient
2. Low latency and high throughput
3. Flexible number format
4. „Easy“ support of new DL architectures

**Cons:**
1. Difficult to program
2. Limited DL software support

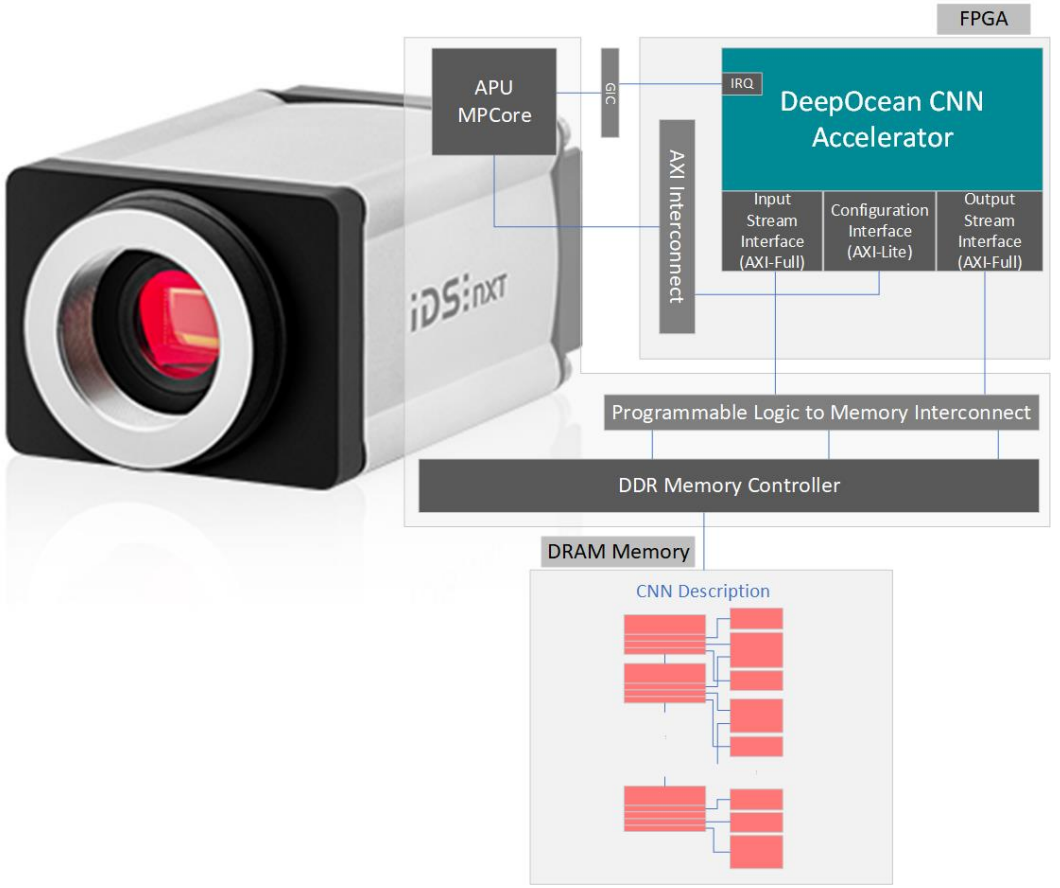
# Benefits of using FPGA technology for DL acceleration

- **Customizable**
- **Optimizable** for specific types of architectures, CNNs
  - reduces power requirements
  - higher performance
- **Longer lifespan** → 2-5 times that of GPUs
- **More resistant** to rugged settings and environmental factors
- **Reconfigurable**
  - ideal when algorithms change frequently,
  - clear advantage over ASIC solutions




# IDS FPGA implementation strategy

- **ONE**, **UNIVERSAL** architecture
- **Any CNN** can run on developed CNN accelerator
- No “Difficult to Program” issue
- Great **flexibility**
- Support of **newest DL algorithms**
- Good DL software **support**



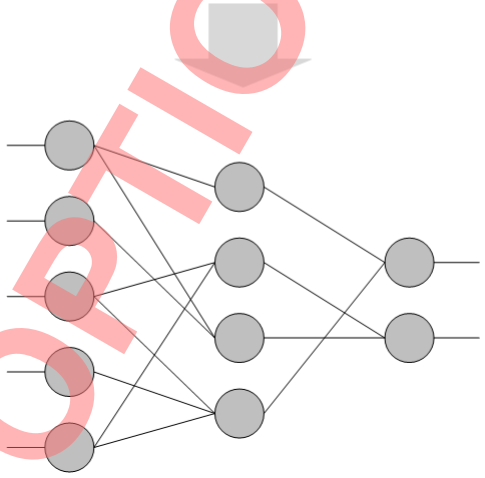
# CNN implementation flow using „Deep Ocean” accelerator

**STEP 1:**

Prune selected CNN

(DeepCompressor Tool)

OPTIONAL

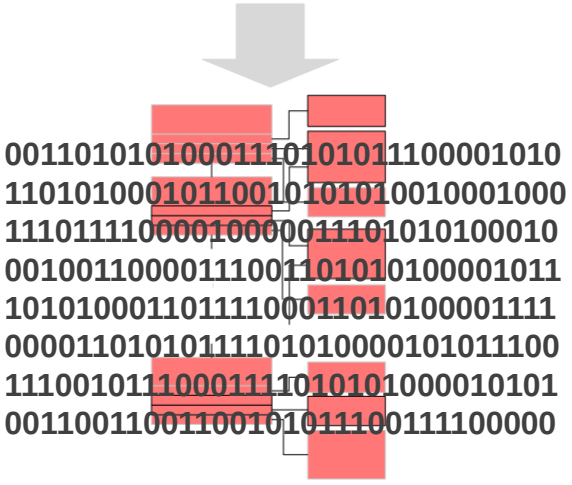


**STEP 2:**

Translate pruned CNN

1. Select number representation
2. Create binary description

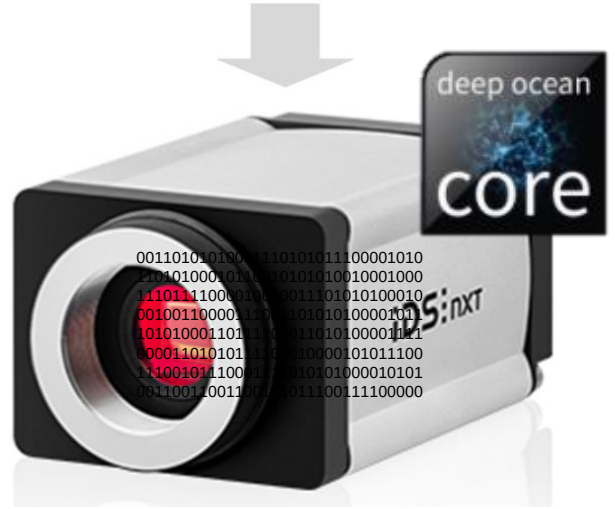
(DeepTranslator Tool)




**STEP 3:**

Run the CNN on Camera

1. Download CNN binary description to camera
2. Configure and start DeepOcean CNN accelerator

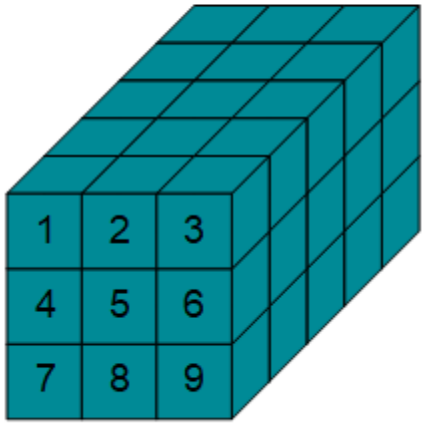



It's so easy!

# Pruning CNNs

**Before pruning**

Every weight value is different from zero

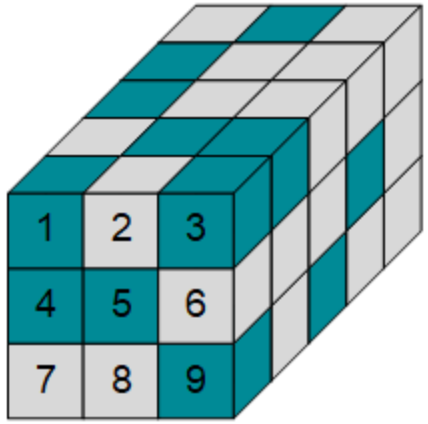



**Original 3x3x5 dense kernel**

**Pruning Algorithm**
Removes specified % of "least important" weights

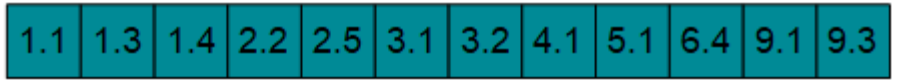
**After pruning**

Some weight values equal to zero (shown in light grey)



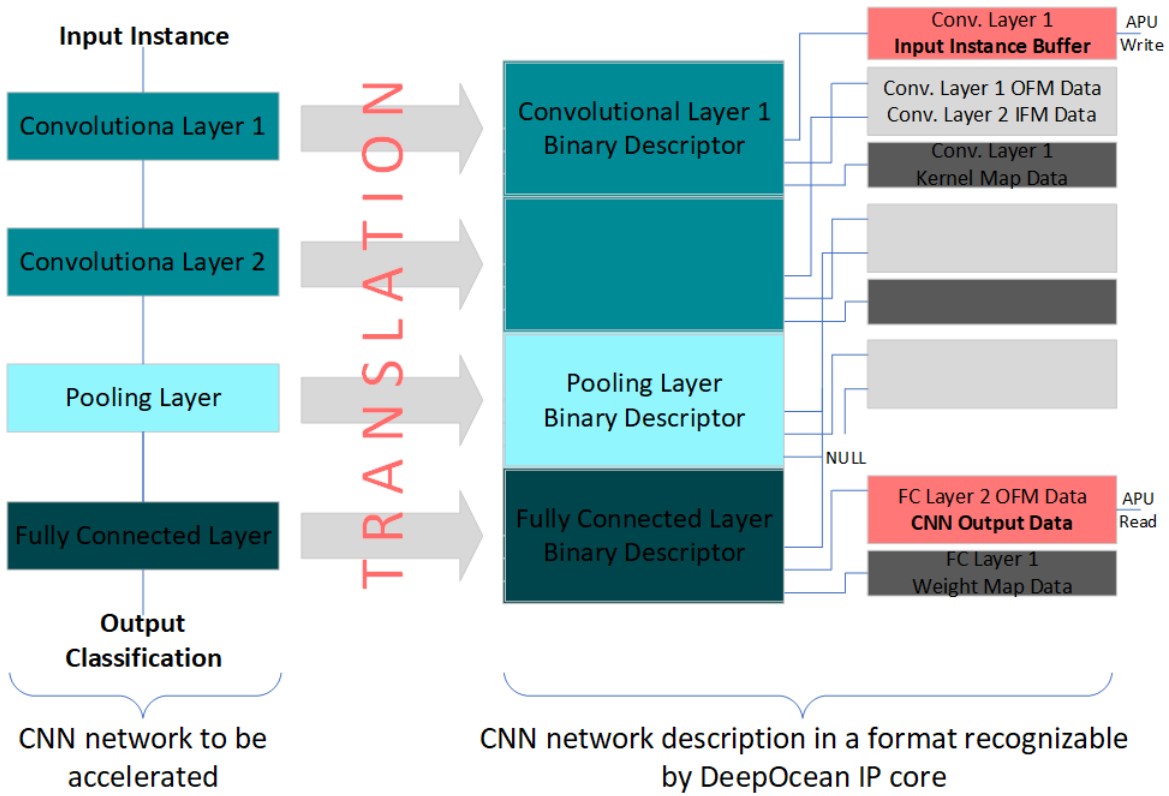

**Sparse 3x3x5 kernel**

A list of non-zero valued weights



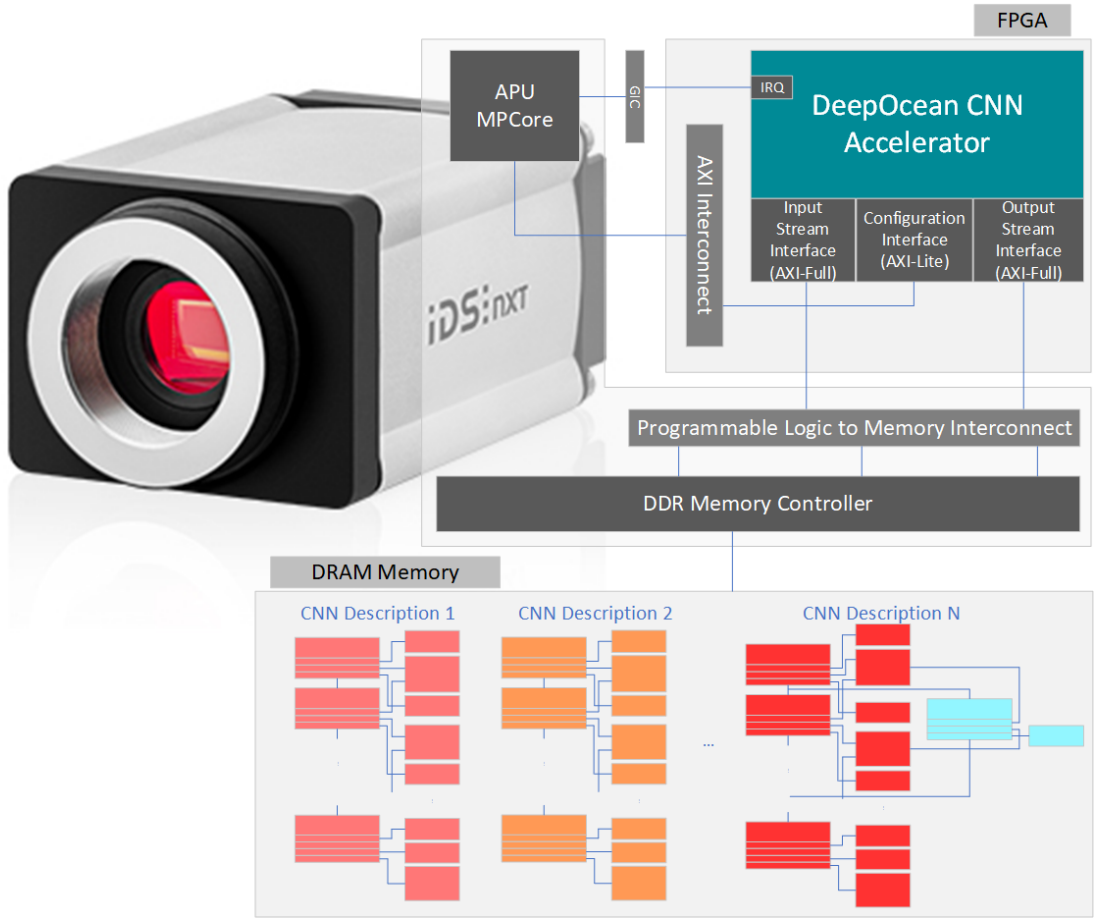
# Preparing CNN to run on „Deep Ocean” accelerator

- **Translate** CNN network model into a binary format
- Determine **optimal** fixed-point number representation
- Floating-point to fixed-point **conversion**
- **Compress** sparse kernels



# Benefits of using linked list CNN representation

- Extremely **easy to change** CNN network
- **Dynamic Switching** between different CNN networks is **extremely fast**
- Enables changing CNN network topology **„on-the-fly“**



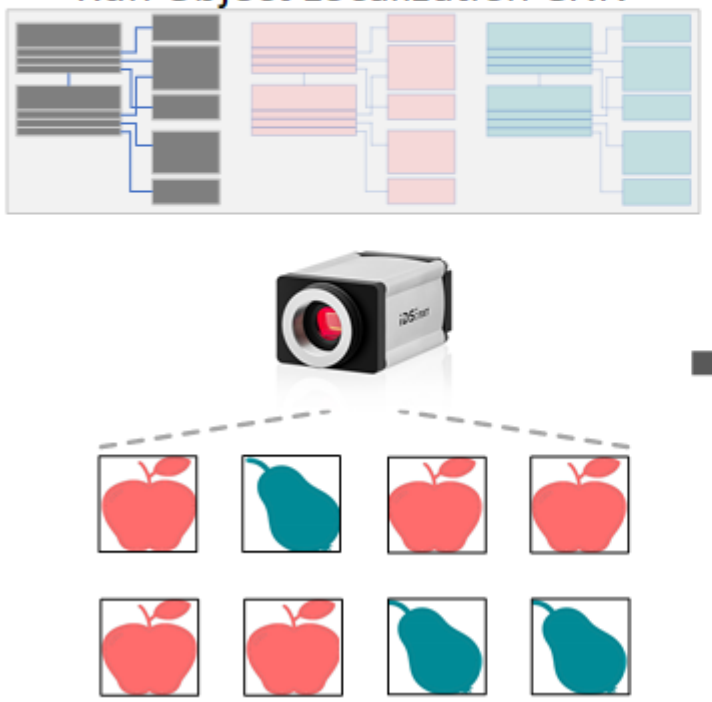
# Advantages of dynamic CNN switching

STEP 1:
Object Localization/Tracking
+ Classification
Run Object Localization CNN

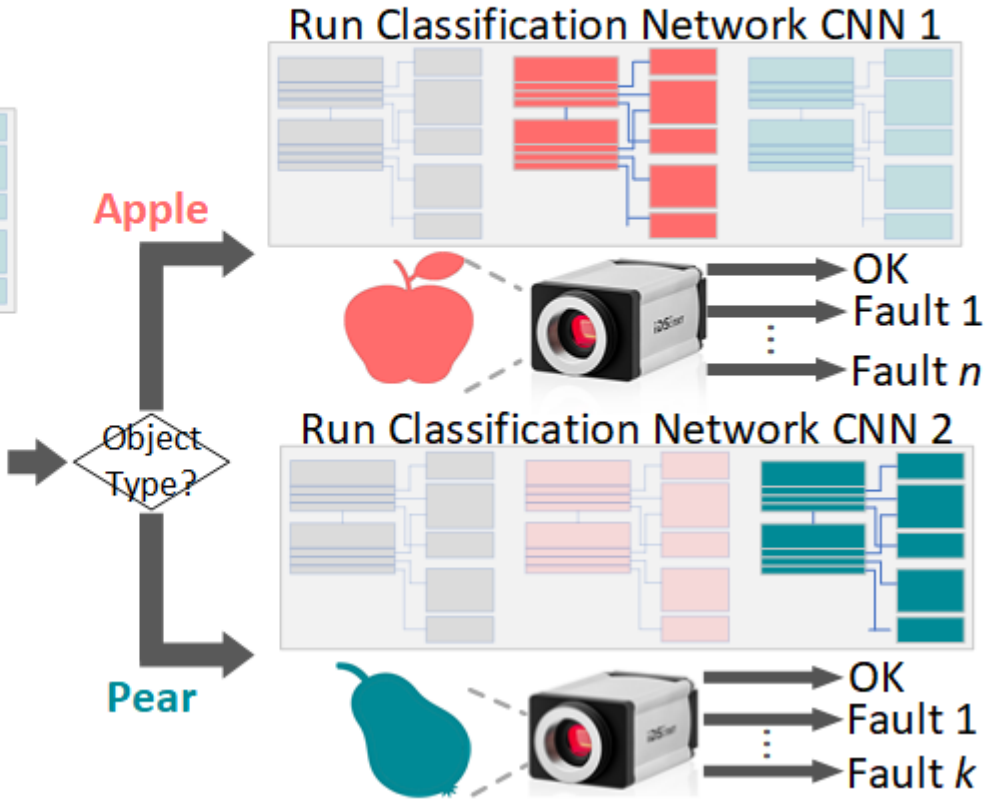

To Difficult to Train!

**STEP 1:**
Object Localization/Tracking

Run Object Localization CNN



**STEP 2:**
Conditional Object Classification



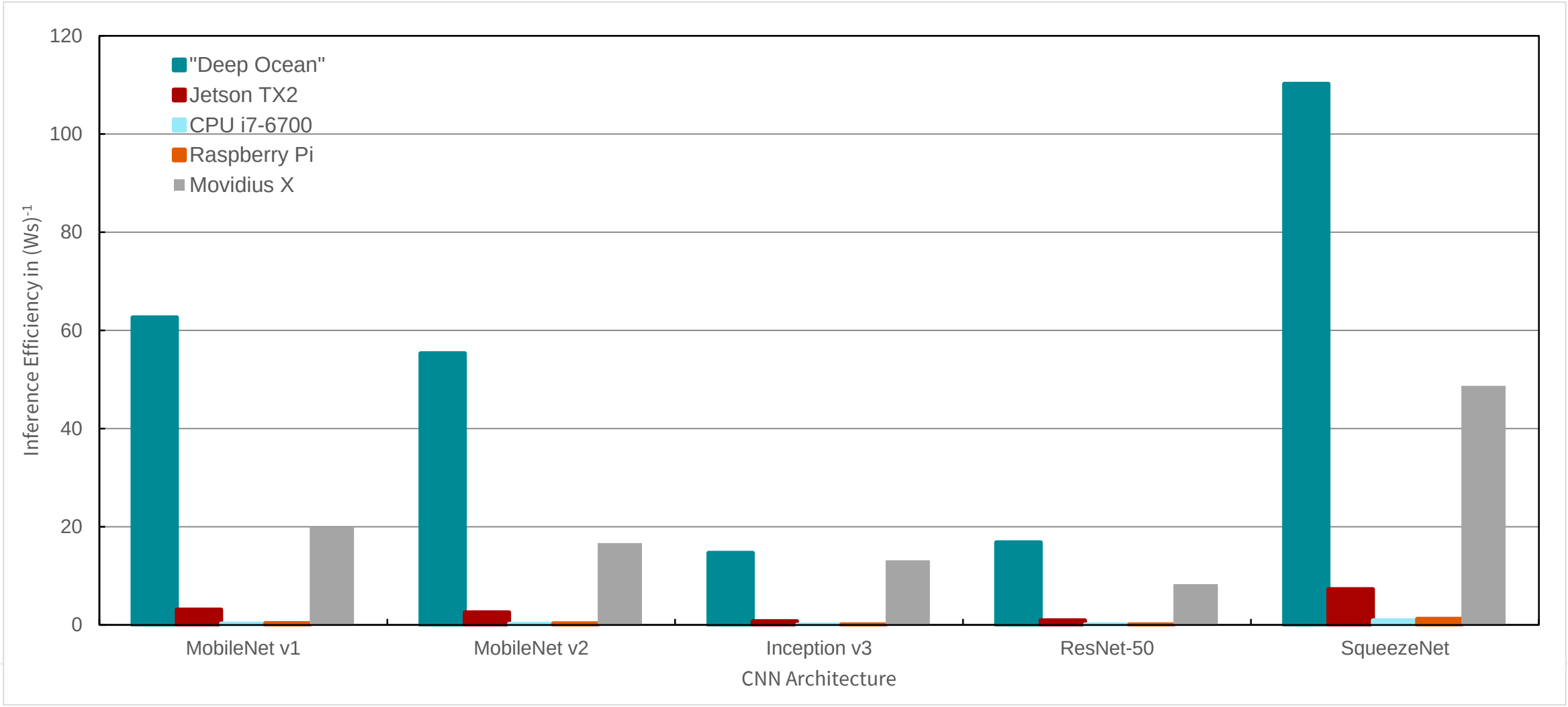
# „Deep ocean” inference latency and throughput

- Target FPGA device Xilinx ZCU3
- Operating at 245 MHz
- Using 64 Compute Cores

| CNN Network | Input Image Size | Latency [ms] | Frame Rate [fps] |
|---|---|---|---|
| MobileNet v1 | 224x224 | 26.16 | 38.23 |
| MobileNet v2 | 224x224 | 29.63 | 33.75 |
| MobileNet v3 large | 224x224 | 34.97 | 28.60 |
| MobileNet v1 SE | 224x224 | 40.1 | 24.94 |
| SqueezeNet | 224x224 | 14.87 | 67.25 |
| EfficientNet B0 | 224x224 | 51.79 | 19.31 |
| MnasNet | 224x224 | 33.43 | 29.91 |
| Inception v3 | 299x299 | 111.62 | 8.96 |
| ResNet-18 | 224x224 | 48.62 | 20.57 |
| ResNet-34 | 224x224 | 61.02 | 16.39 |
| ResNet-50 | 224x224 | 97.39 | 10.27 |
| MobileNet v1 SSD | 300x300 | 48.20 | 20.75 |
| MobileNet v2 SSD | 300x300 | 50.65 | 19.74 |

Actual performance of IDS NXT camera can differ from numbers presented above

# „Deep ocean” power efficiency



# THANK YOU!

**IDS Imaging Development Systems GmbH**

Dimbacher Str. 6-8, 74182 Obersulm I Tel.: +49 7134 96196-0 I www.ids-imaging.com

# Copyright Notice



www.tinyML.org