tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Evolutionary Needs of TinyML"

Liang Shen - Qualcomm

January 13, 2022

www.tinyML.org

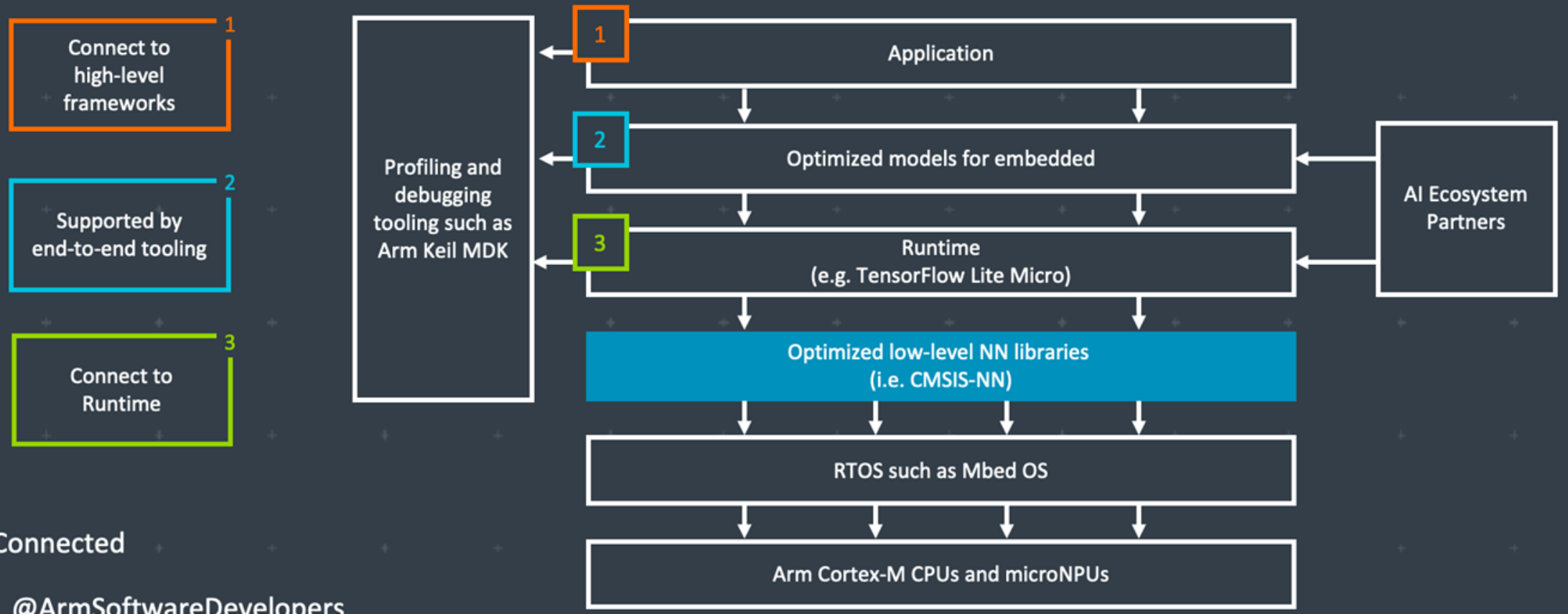
tinyML Talks Strategic Partners

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

TINY ML TALKS webcast

Executive Strategic Partners

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

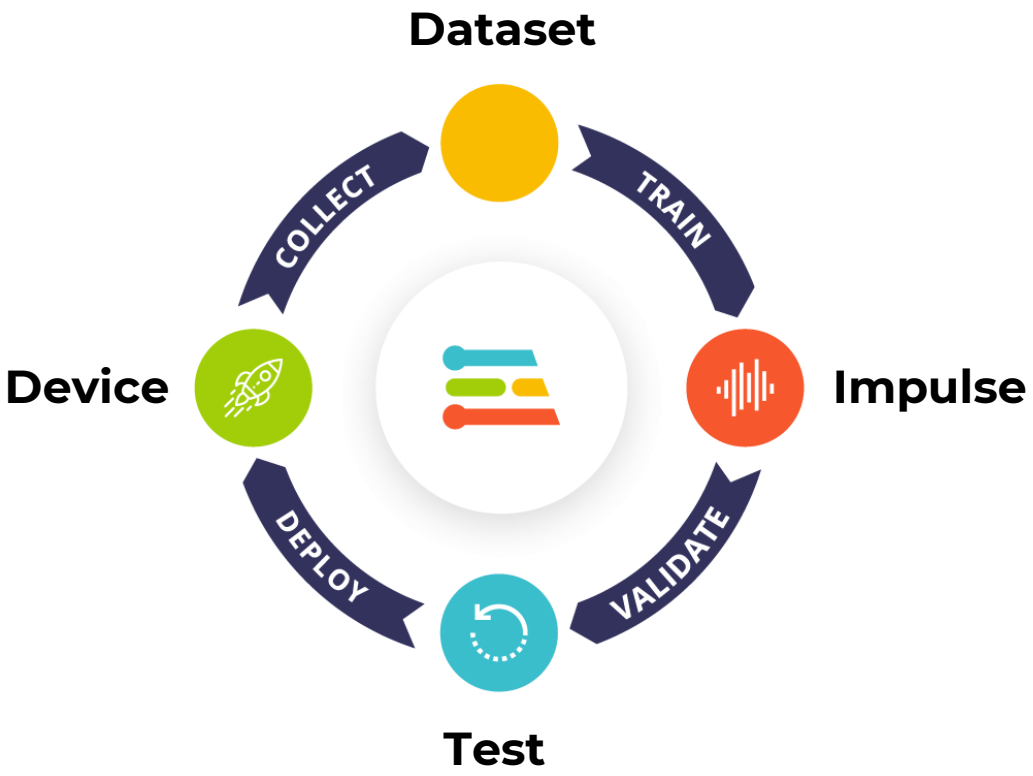
@ArmSoftwareDev

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

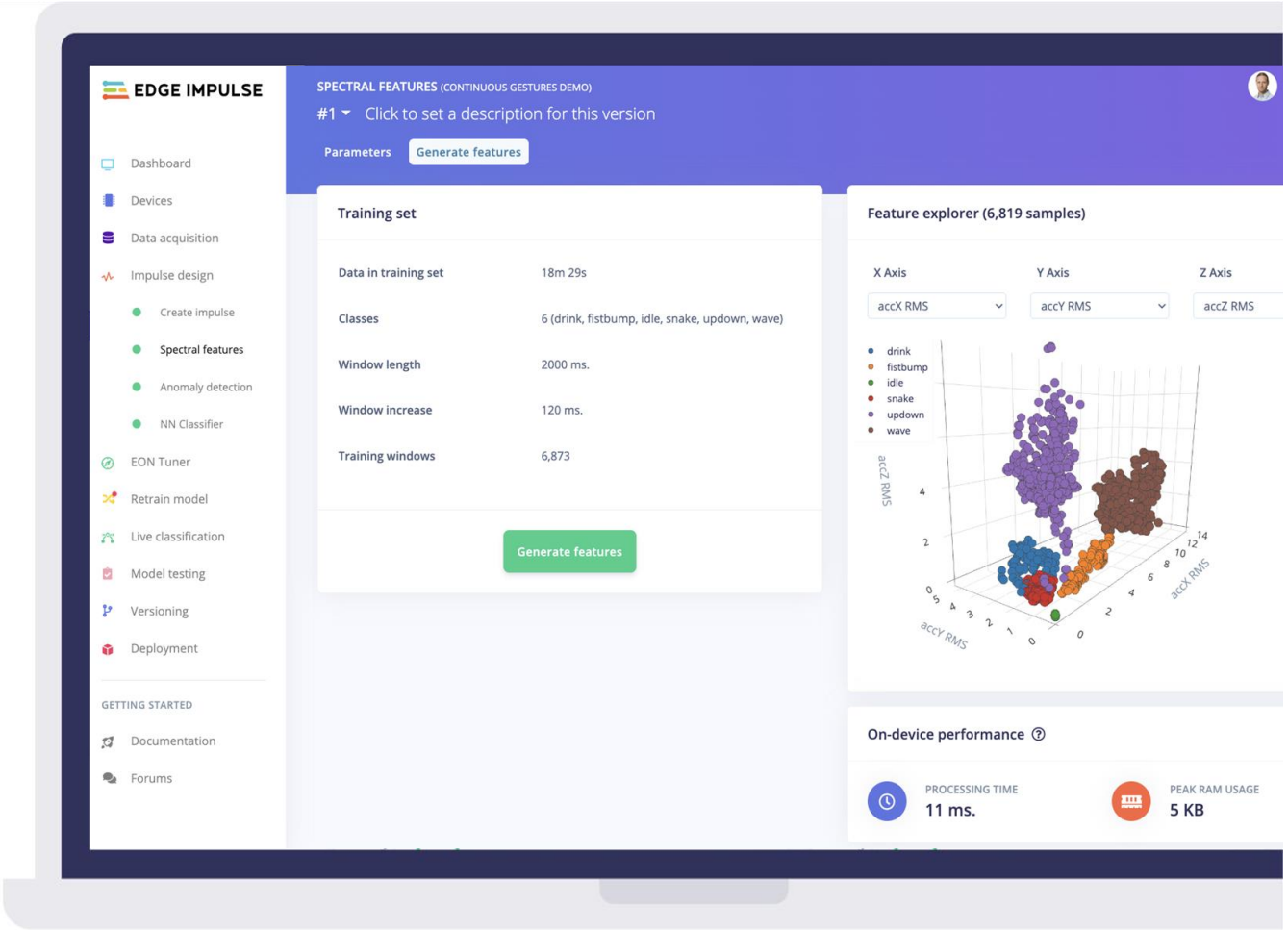
© 2020 Arm Limited (or its affiliates)

arm

EDGE IMPULSE The leading edge ML platform



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

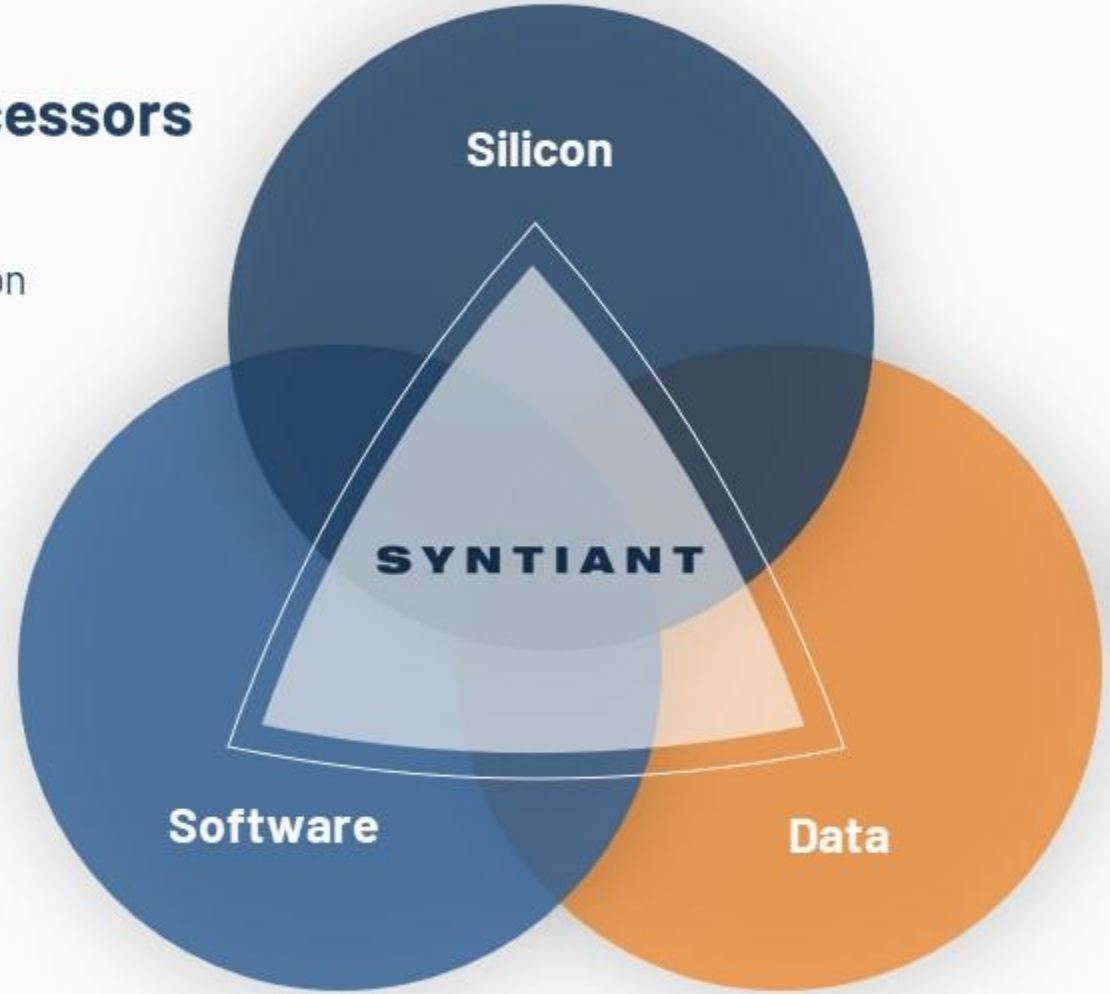
Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

SYNTIANT

End-to-End Deep Learning Solutions for TinyML & Edge AI

Neural Decision Processors

- At-Memory Compute
- Sustained High MAC Utilization
- Native Neural Network Processing



ML Training Pipeline

- Enables Production Quality Deep Learning Deployments

Data Platform

- Reduces Data Collection Time and Cost
- Increases Model Performance

partners@syntiant.com

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

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

TINY ML TALKS webcast

KLIKA·TECH

GLOBAL IOT SOLUTIONS

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

BROAD AND SCALABLE EDGE COMPUTING PORTFOLIO

Microcontrollers & Microprocessors

Arm® Core

RENESAS RA
Arm® Cortex®-M 32-bit MCUs
Arm ecosystem, Advanced security, Intelligent IoT

RENESAS RZ
Arm®-based High-end 32 & 64-bit MPUs
High-resolution HMI, Industrial network & real-time control

RENESAS RE
Arm® Cortex®-M0+ Ultra-low Power 32-bit MCUs
Innovative process tech (SOTB), Energy harvesting

Renesas Synergy™
Arm®-based 32-bit MCUs for Qualified Platform
Qualified software and tools

Renesas Core

RENESAS RL78
Ultra-low Energy 8 & 16-bit MCUs
Bluetooth® Low Energy, SubGHz, LoRa®-based Solutions

RENESAS RX
High Power Efficiently 32-bit MCUs
Motor control, Capacitive touch, Functional safety, GUI

RENESAS RH850
40nm/28nm process Automotive 32-bit MCUs
Rich functional safety and embedded security features

Core technologies

AI

A broad set of high-power and energy-efficient embedded processors

Security & Safety

Comprehensive technology and support that meet the industry's stringent standards

Digital & Analog & Power Solution

Winning Combinations that combine our complementary product portfolios

Cloud Native

Cross-platforms working with partners in different verticals and organizations

© 2021 Renesas Electronics Corporation. All rights reserved.

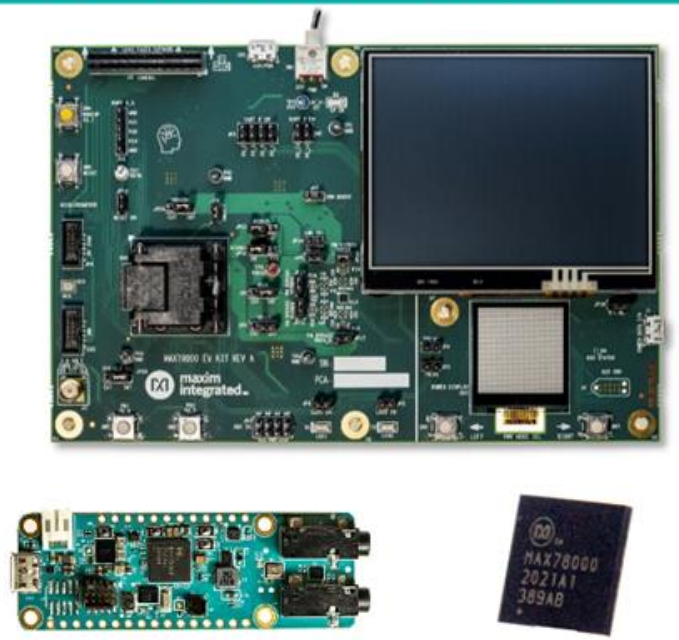
BIG IDEAS FOR EVERY SPACE RENESAS

TINY ML TALKS webcast

Gold Strategic Partners

Maxim Integrated: Enabling Edge Intelligence

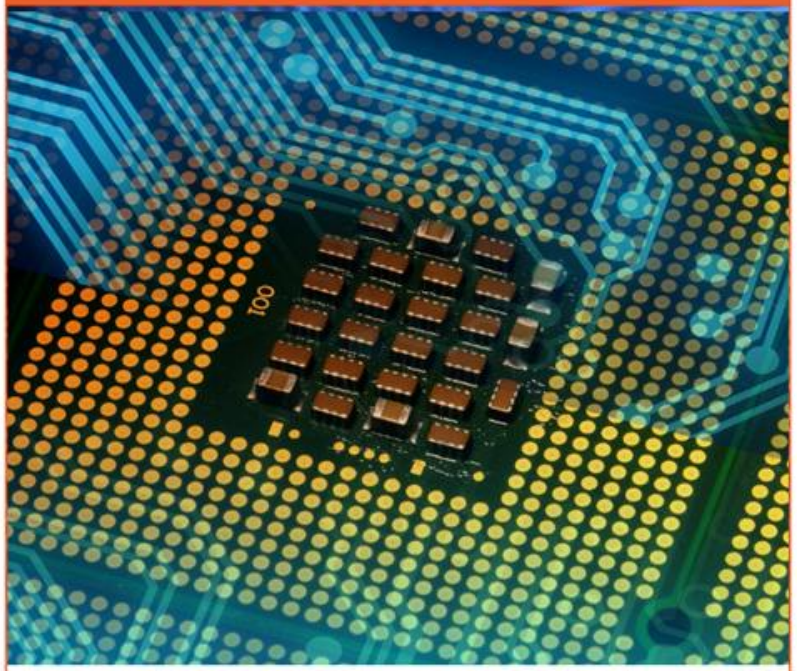
Advanced AI Acceleration IC



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

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

TINY ML TALKS webcast

seeed studio

The IoT Hardware Enabler

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.

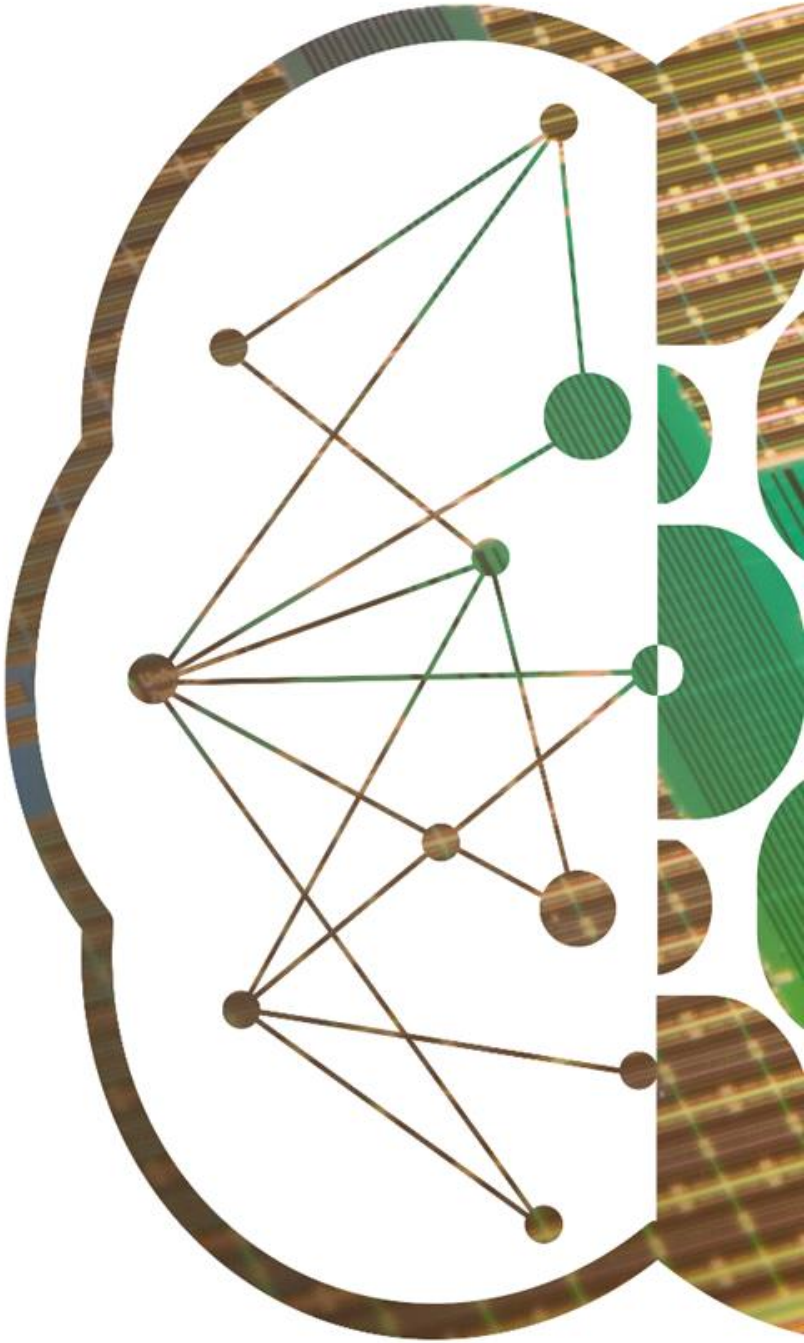
sensiml.com

TINY ML TALKS webcast

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

Silver Strategic Partners

AONdevIces

emza visual sense

GREENWAVES TECHNOLOGIES

Groovety Inc.

HOTG

imagimob

Qeexo

PROPHESEE METAVISION FOR MACHINES

REXEN technology

SAP

SA STREAM ANALYZE

tinyML Summit 2022

Miniature dreams can come true…

March 28-30, 2022

Hyatt Regency San Francisco Airport

https://www.tinyml.org/event/summit-2022/

The Best Product of the Year and the Best Innovation of the Year awards are open for nominations between ***November 15 and February 28.***

tinyML Research Symposium 2022

March 28, 2022

https://www.tinyml.org/event/research-symposium-2022

More sponsorships are available: sponsorships@tinyML.org

Join Growing tinyML Communities:

Meetup

7.7k members in
42 Groups in 33 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

2.5k members
&
4.3k followers

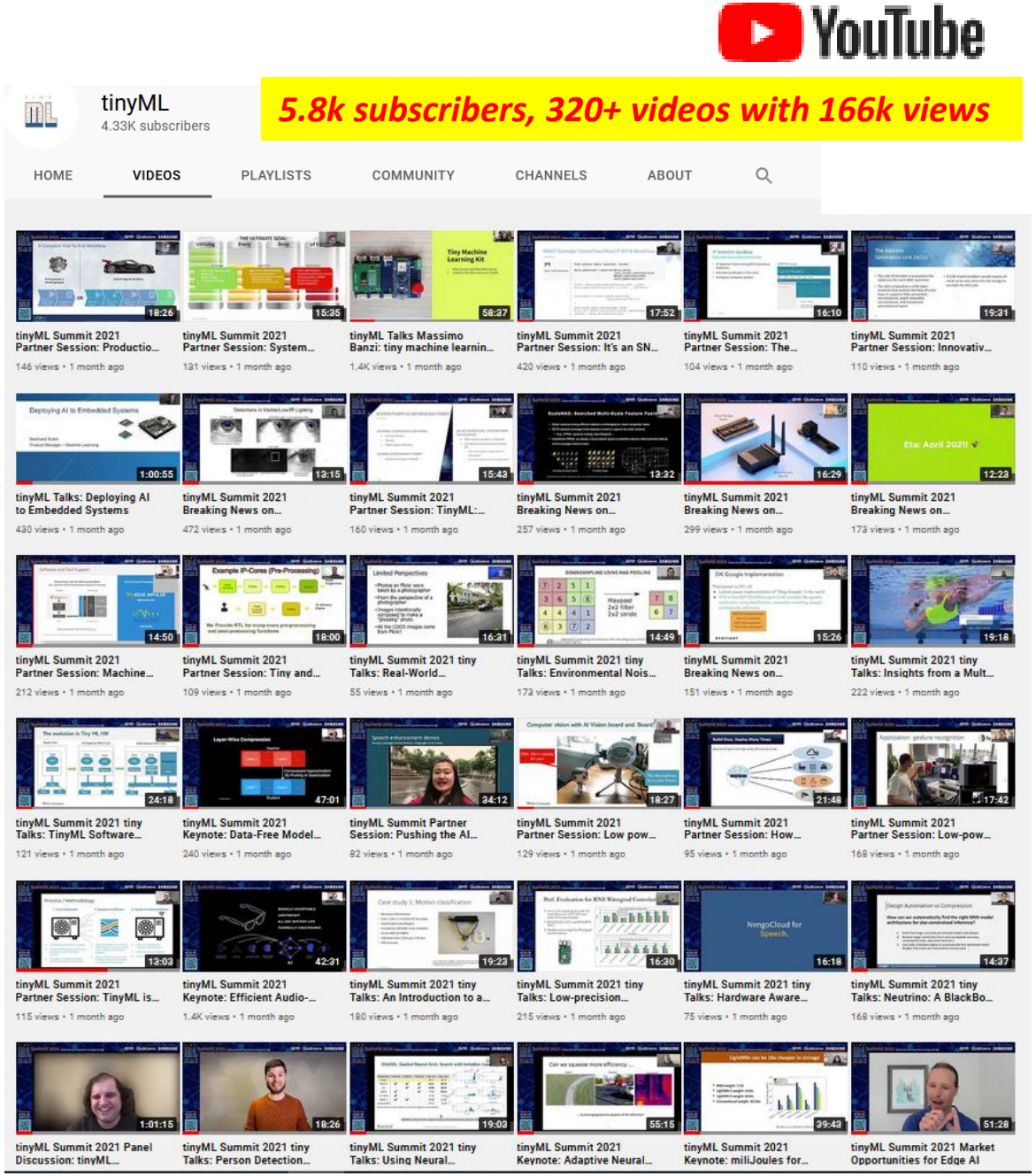
The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML

Subscribe to tinyML YouTube Channel for updates and notifications *(including this video)*

www.youtube.com/tinyML



Next tinyML Talks

Date	Presenter	Topic / Title
Tuesday, January 18	Ashutosh Pandey, Infineon Technologies	Exploring techniques to build efficient and robust TinyML deployments

Webcast start time is 8:00 am Pacific time

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

TINY ML TALKS webcast

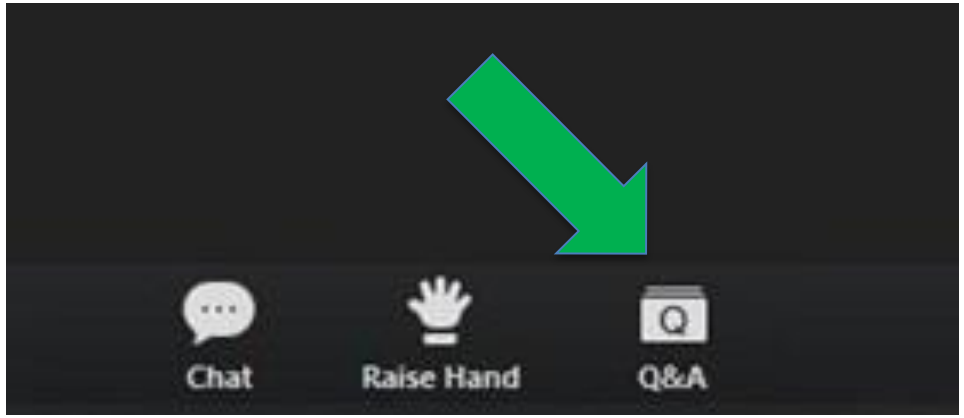
Reminders

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

Liang Shen

With early training in the fields of Biomedical Engineering and Artificial Intelligence & Pattern Recognition, pursuing dream to make IoT becoming Wisdom-of-Things (WoT).

Jan 13, 2022

Evolutionary Needs of TinyML

Liang Shen
Qualcomm

Abstract

During the past decade, Deep Learning based AI technology not only becomes the predominant solutions for existing or new problems, also almost instantly deployed to various smart devices. In this talk, we start with a brief review on how power-efficient AI engine helped this new AI wave and effectively enabled billions of battery-powered devices; then, we touch the new trend: always-on or long-continuous-run AI use cases, which require optimal minimum power solution. We discuss some details of ultra-low-power AI solution and how it offers the improved quality for targeted use cases. With continuous evolution of new intelligent algorithms, this talk concludes with on-going challenges and some potential directions.

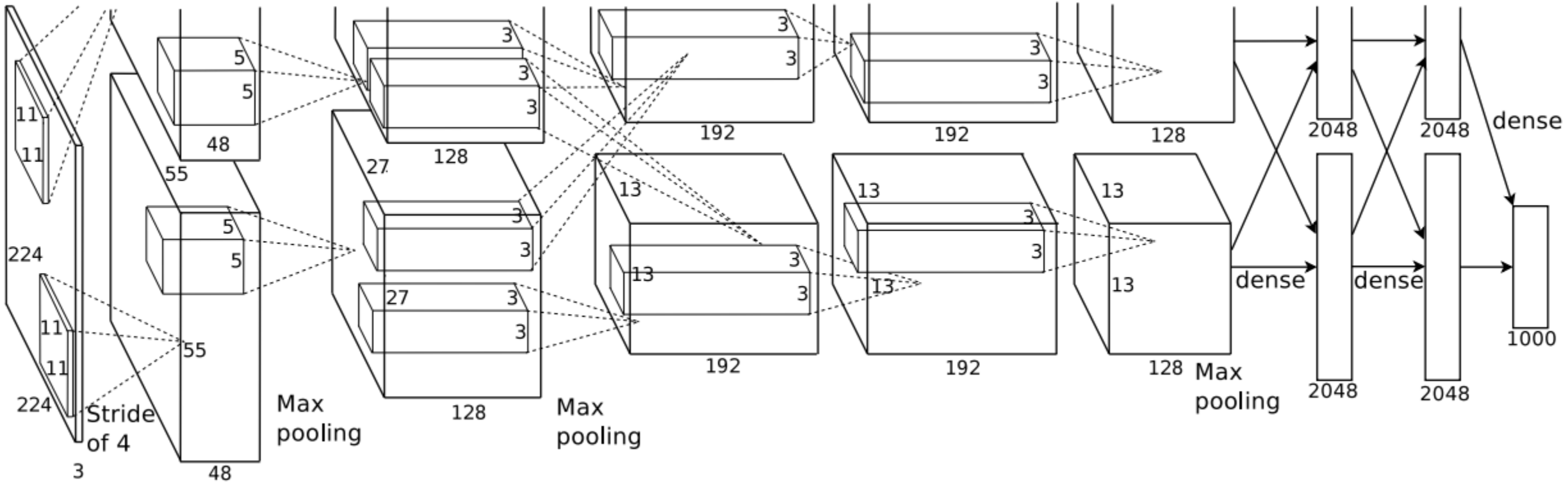
Outline

- Efficient On-Device AI
- Ultra Low Power AI
- Challenges and Directions

Outline

- Efficient On-Device AI
- Ultra Low Power AI
- Challenges and Directions

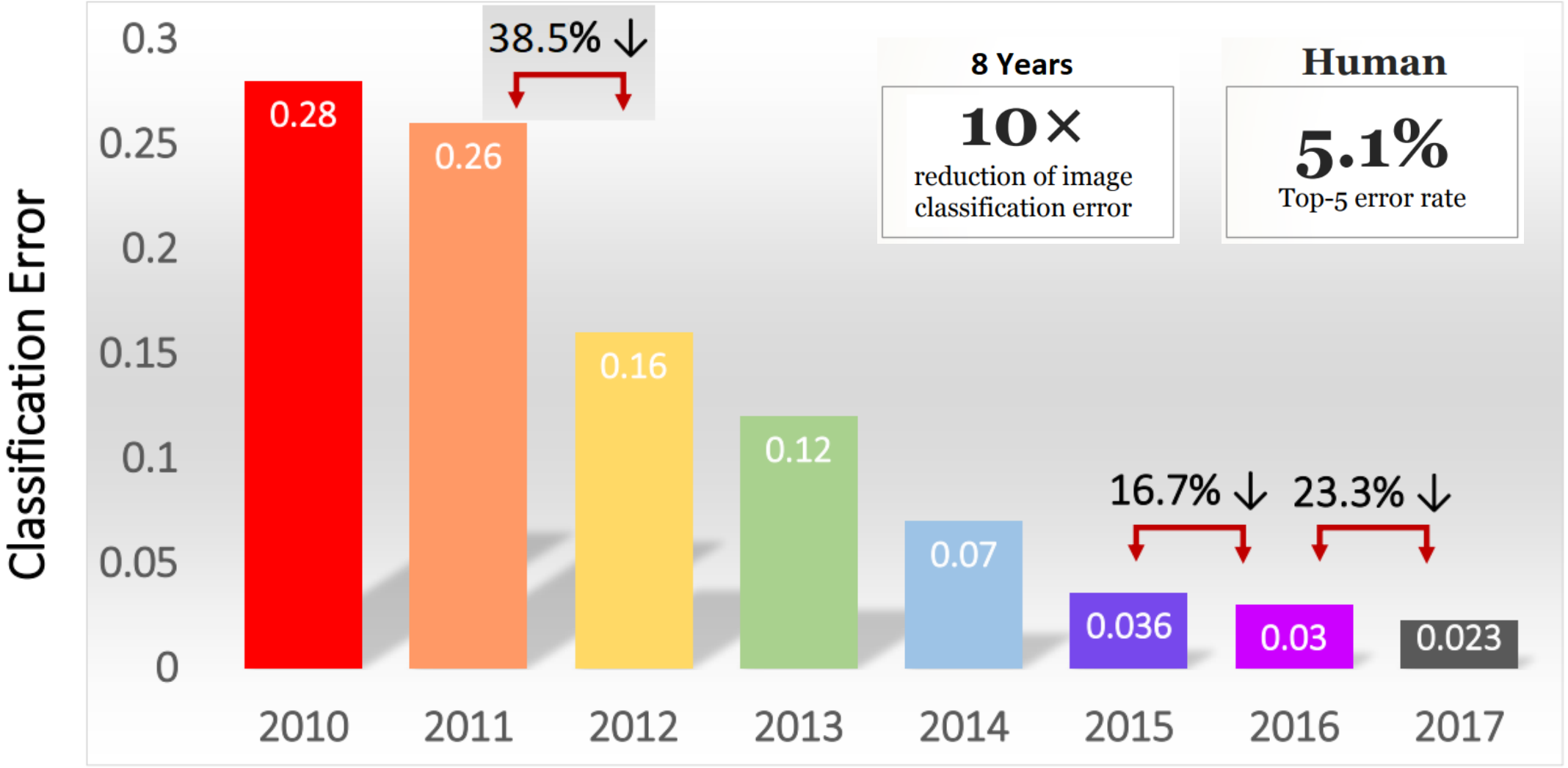
Triggering Point of the Current AI Wave



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

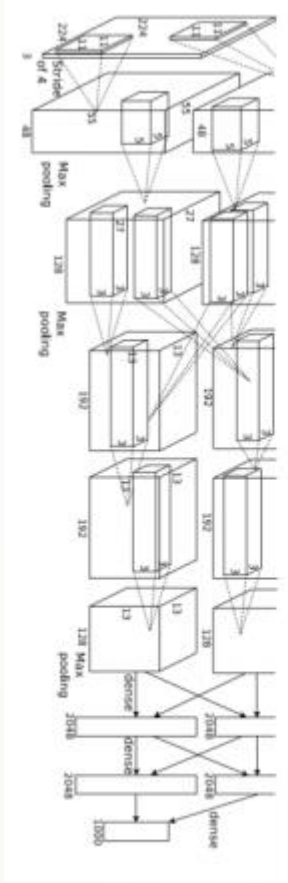
NIPS 2012



Large Scale Visual Recognition Challenge (2010 – 2017)

http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf
http://image-net.org/challenges/talks_2017/ILSVRC2017_overview.pdf

AlexNet
(232MB, 724M MAC)

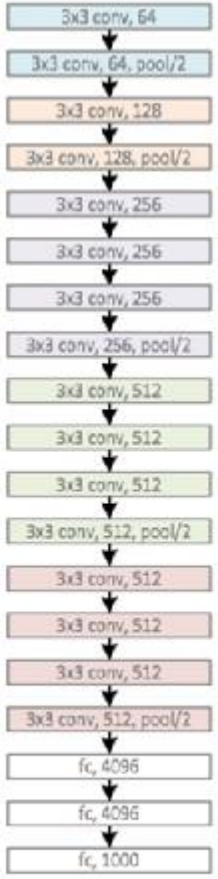


[Krizhevsky et al. NIPS 2012]

GoogLeNet-V1
(26MB, 1582M MAC)

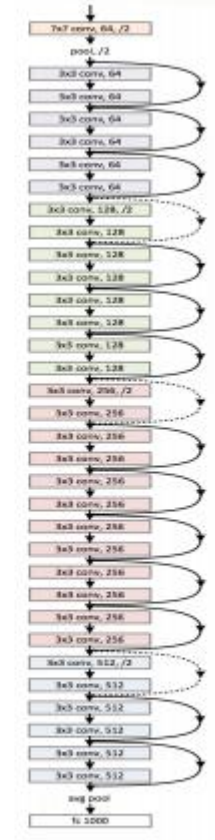
[Szegedy et al. CVPR 2015]

VGG-19
(547MB, 19632M MAC)



[Simonyan & Zisserman, ICLR 2015]

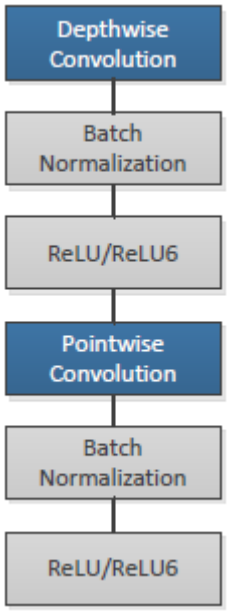
ResNet-152
(222MB, 11280M MAC)



[He et al. CVPR 2016]

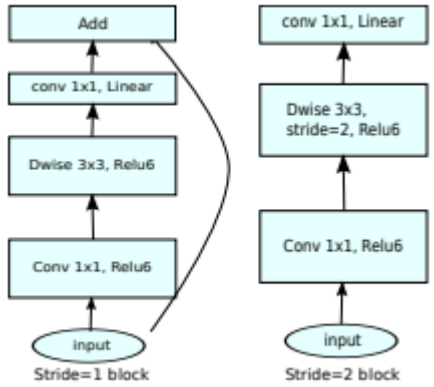
Good Accuracy with Large Model Size and High Computational Cost

MobileNetV1
(4.2MB, 575M MAC)



[Howard et al. 2017]

MobileNetV2
(3.4MB, 300M MAC)



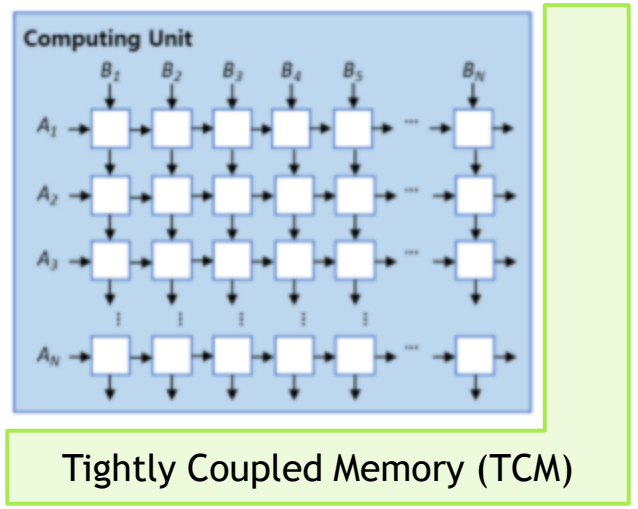
[Sandler et al. 2018]

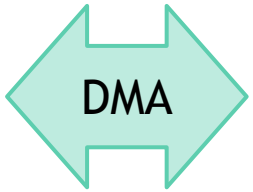
Network	Top 1	Params	MAdds	CPU
MobileNetV1	70.6	4.2M	575M	113ms
ShuffleNet (1.5)	71.5	**3.4M**	292M	-
ShuffleNet (x2)	73.7	5.4M	524M	-
NasNet-A	74.0	5.3M	564M	183ms
MobileNetV2	**72.0**	**3.4M**	**300M**	**75ms**
MobileNetV2 (1.4)	**74.7**	6.9M	585M	**143ms**

Network	Snapdragon SM8150 (2017-2018)
MobileNetV1	1.78ms
MobileNetV2	2.44ms
InceptionV3	9.71ms
InceptionV4	19.89ms
ResNet34	5.78ms
ResNet50	7.98ms
ResNet101	15.99ms
VGG16	28.13ms
VGG19	33.19ms

Lightweight Model Architecture for On-Device AI

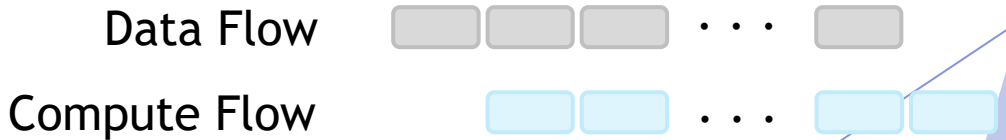
CPU – GPU – DSP - … → NPU: HTA, HTP





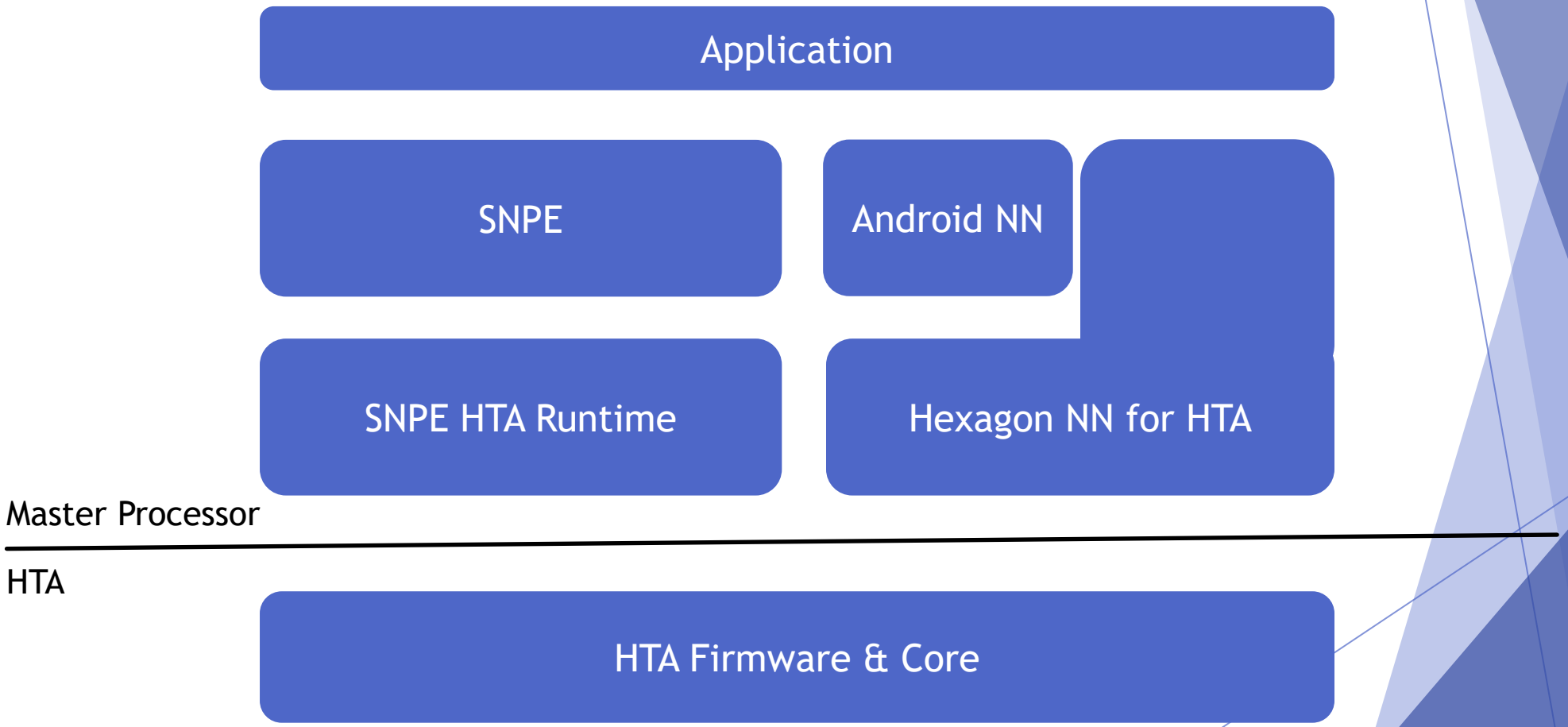
SRAM
or
DDR

- Parallel 8/16bit Fixed-Point Processing Units
- Parallel Processing Workflow



Qualcomm® HTA Software Architecture

Dedicated AI/DNN hardware accelerator



Outline

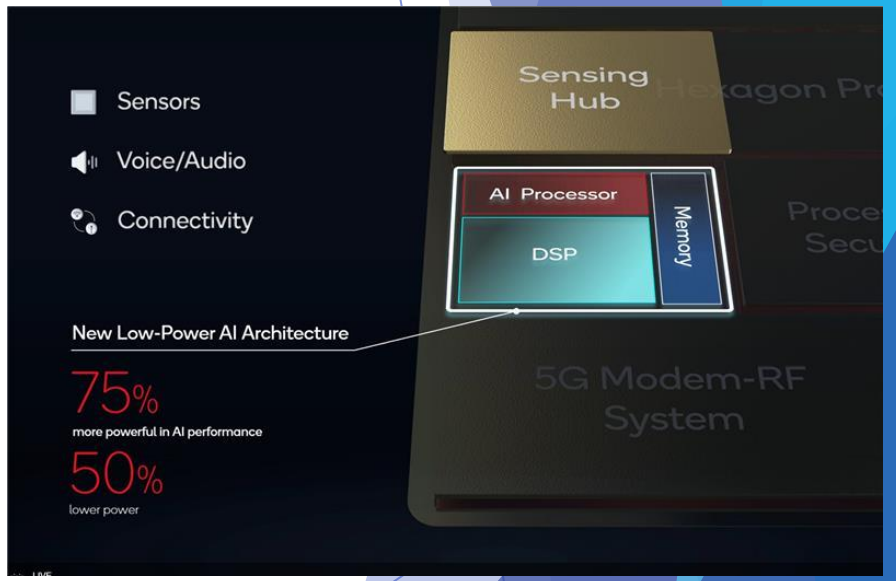
- Efficient On-Device AI
- Ultra Low Power AI
- Challenges and Directions

Video Super-Resolution: 360p to 1080p

Although 5ms model inferencing time could be achieved at Turbo mode, we could run at LowSVS for 22ms of model inferencing time with power consumption below 100mA.

Always-On or Long-Duration Use Cases: AI-based solution makes them possible and/or better

- Always-on speech recognition, always-on sensor action recognition, and always-on camera, etc
- Contextual awareness and audio enhancement: noise cancellation and acoustic scene Recognition, etc.



Focus: Performance → Optimal Power at Target Performance

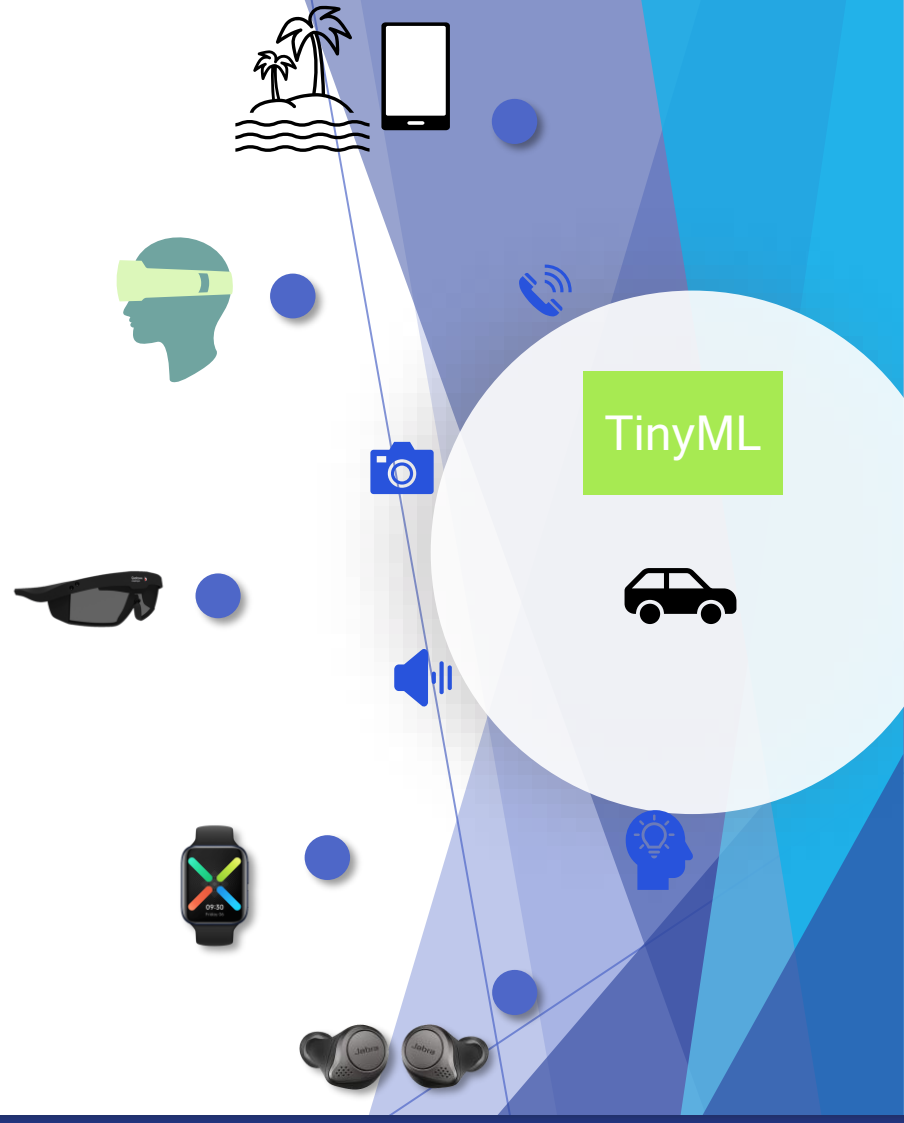
TinyML Requirements

100 - 1000x Differences

- Compute Capability: 0.1–1 vs 1-100 TOPS
- Memory Availability: 0.1–2 vs 2–200 MB
- Power Consumption: 1–20 vs 50-2000 mA

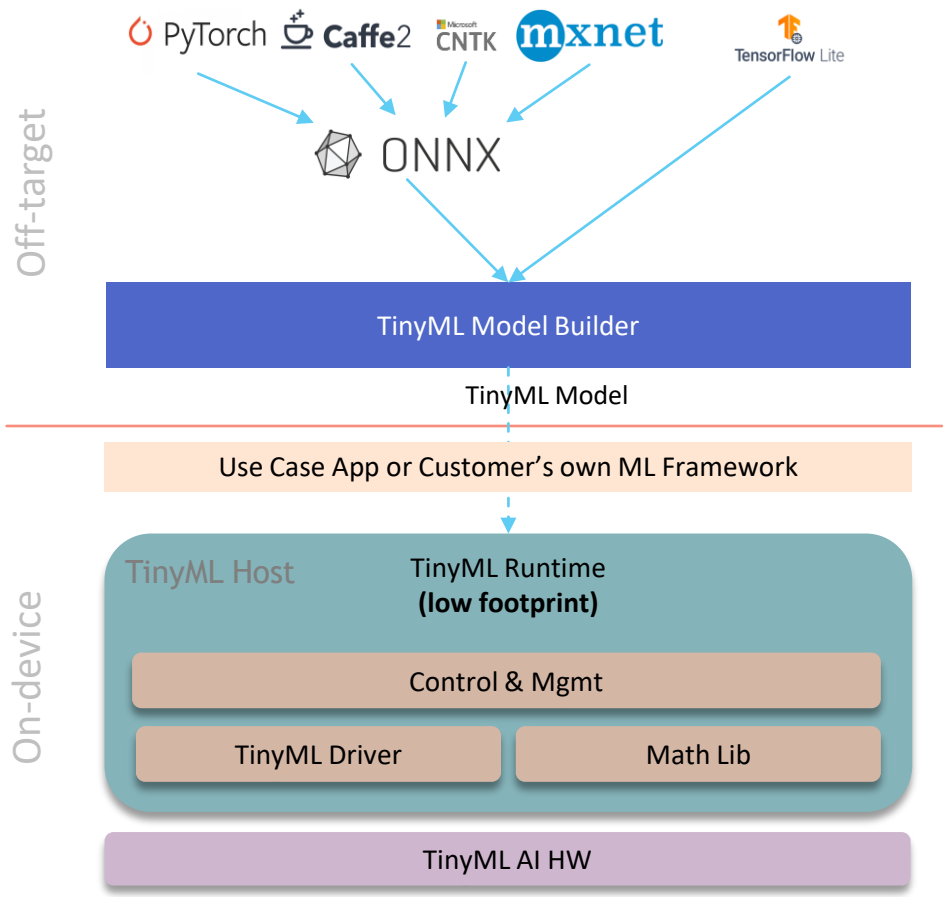
TinyML Solution

- Lightweight ML Models
- Low-overhead Inference Framework
- Efficiency in KB memory



Challenges for Broader Benefits

Optimal Maximum Performance → Optimal Minimum Power



Model Builder

Convert to a proprietary representation which can be processed more efficiently on devices:

- Model Quantization
- Graph Transformation: general Op fusion and specific target Op mapping
- HW formation: weights reordering, tiling, compression, etc.
- Address/padding alignments
- Memory planning
- Serialization and model generation

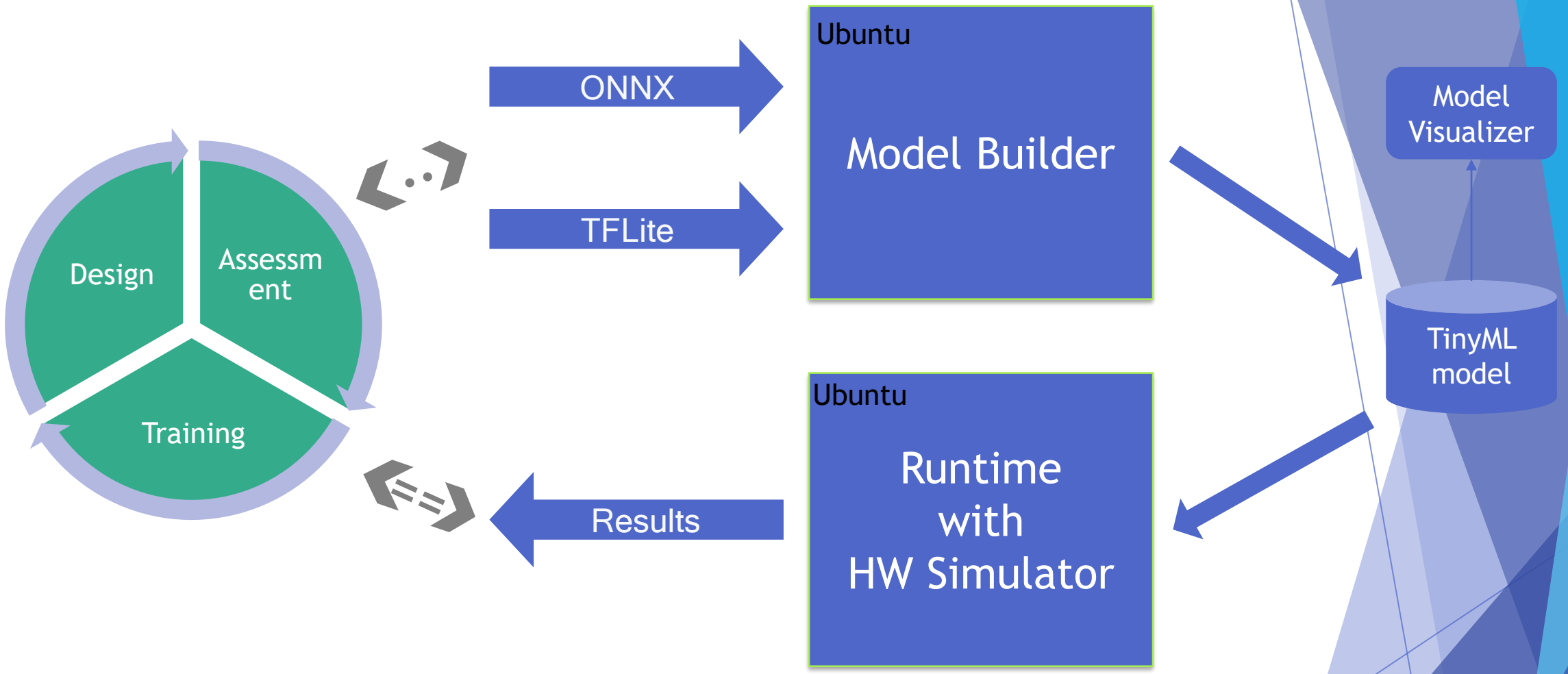
Runtime

Execute model with minimal latency, as well as optimized memory footprint and power consumption:

- init() / deinit()
- preapply()
- apply()
- set_property() / get_property()
- execute()

Ultra Low-Power AI Solution

https://developer.qualcomm.com/software/hexagon-dsp-sdk/tools



TinyML SDK: Complete Off-Target Tools for Model Verification

Outline

- Efficient On-Device AI
- Ultra Low Power AI
- Challenges and Directions

Challenges and Directions

Tiny but generic and customizable

- Framework: general TinyML customizable framework
- Tools: intuitive and easy to use

Tiny without quality compromizations

- Model Size: HW-Friendly training, HW-aware NAS, sparsity, low-bit-precision
- Activation Size: compression, execution flow optimization

Tiny and learning

- Refinement
- Personalization

The brain creates a predictive model. This just means that the brain continuously predicts what its inputs will be. Prediction isn't something that the brain does every now and then; it is an intrinsic property that never stops, and it serves an essential role in learning. When the brain's predictions are verified, that means the brain's model of the world is accurate. A mis-prediction causes you to attend to the error and update the model.

Jeff Hawkins: "A Thousand Brains: A New Theory of Intelligence"

Thank you

Copyright Notice

This multimedia file is copyright © 2022 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org

Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. 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