


tinyML[®] Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area



www.tinyML.org



The background features a light blue gradient with a subtle hexagonal grid. In the upper left, a cluster of small white triangles forms a larger hexagonal shape. Scattered throughout are several 3D cubes in various shades of blue, cyan, and red, some with bright highlights. Faint mathematical formulas are visible: $F(s) = -e^{-sT} \left(\frac{\alpha}{s+\alpha} \right)$ on the left, $f_m(t) = 1 - e^{-\beta t}$ on the right, and $F(s) = e^{-s}$ on the far right. A red horizontal bar is positioned above the title.

Dissecting a Low-Power AI/ML Edge Application: Noise Suppression

Raj Pawate, Group Director, Cadence
TinyML Summit 2022
March 2022

Zoom Fatigue Is Real! ...a WFH Side Effect



Speech/audio quality is
key contributor to
video-conferencing
fatigue

Stanford University Research

<https://news.stanford.edu/2021/04/13/zoom-fatigue-worse-women/>

Ways to Alleviate Cognitive Overload

- Reduce noise (both stationary and dynamic)
- Focus on speaker of interest
- Make speech more intelligible
- Increase audio bandwidth: wideband (32KHz, 48KHz)

Agenda

- Challenges for implementing noise suppression (NS)
- R&D in NS
- A holistic Cadence® solution for NS that addresses these challenges
 - Tensilica® HiFi 5 DSP coupled with a HWA NNE110
- Performance and energy benefits of solution
- Conclusion

Challenges for NS

1. End users will not tolerate *delays* in conversations
 - ✓ An algorithmic delay of less than 40ms required; ideally less than 5-10ms
2. The chosen NS algorithm and its implementation **cannot** accelerate battery drain
 - ✓ Especially important for wearable, smart phone, ear-bud, and laptop applications
3. How can you rapidly integrate NS with **other** components in a resource-constrained product?
 - ✓ NS is not an end-product by itself; it is a **front-end** to other audio applications such as ASR, codecs, AEC, etc

All of these components still *need to fit within memory and compute cycle budgets of each end-product*

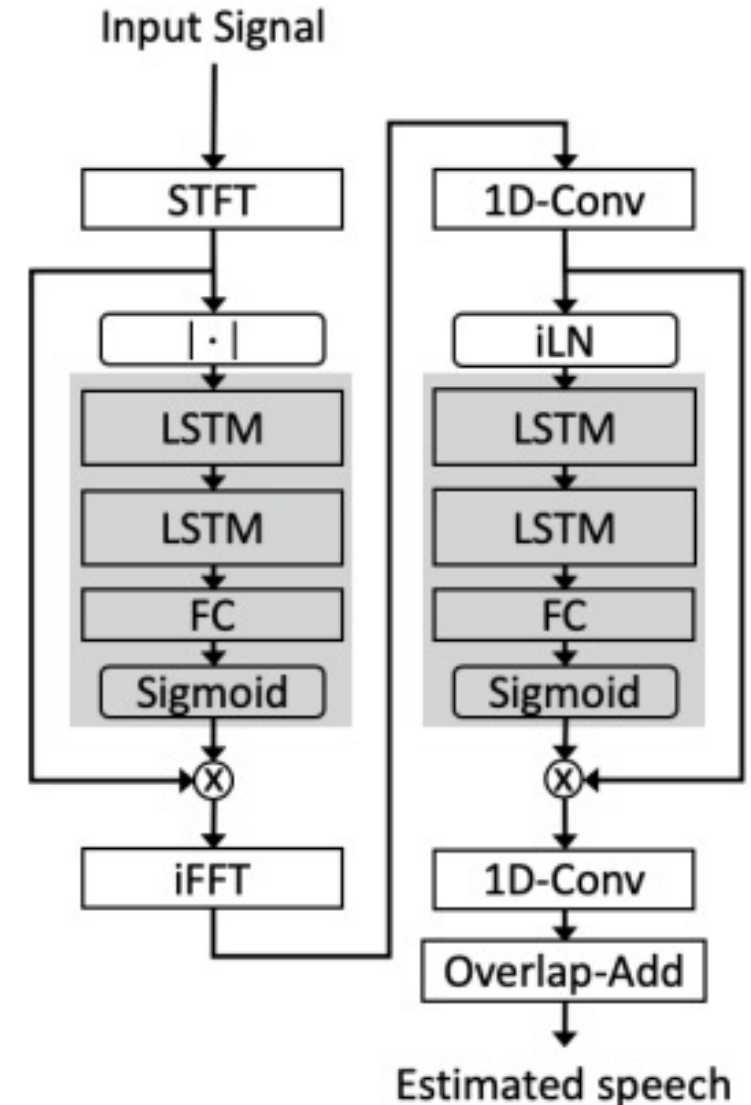
The Rise of ML Algorithms for NS

- Deep Noise Suppression (DNS) Challenge has motivated a lot of R&D
 - ✓ DTLN, DPRNN, TSTNN, ...
- Building block operators typically used are
 - ✓ LSTM for modelling time series, CNN, and others like BiLSTM, Transformer...
- In this presentation, we discuss how Cadence
 - ✓ Created and optimized a hardware-software platform for NS based on these operators while solving these challenges



LSTM and CNN-Based NS NNs

- LSTM-based NS algorithms
 - ✓ Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression by N.L. Westhausen & B.T. Meyer
 - Carl von Ossietzky University, Oldenburg, Germany
 - <https://arxiv.org/pdf/2005.07551.pdf>
- CNN-based NS algorithm
 - ✓ [GitHub - vbelz/Speech-enhancement: Deep learning for audio denoising](#)

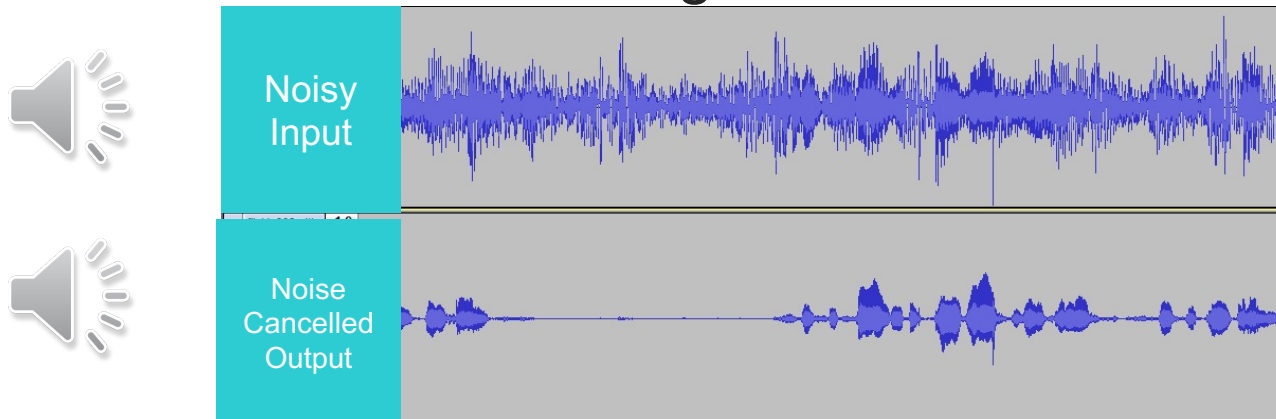


Noisy Input and ML Noise Suppressed Examples

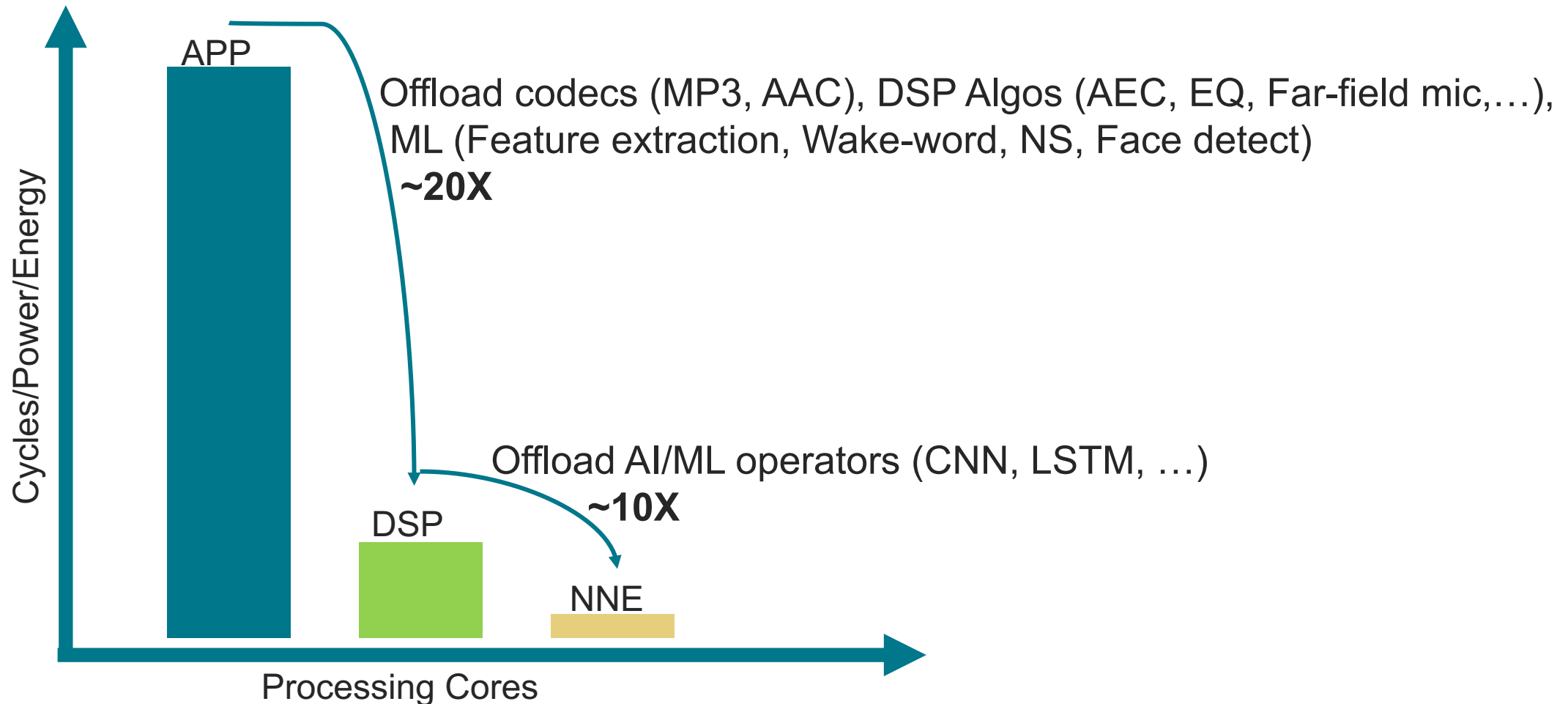
- Case-1: Dog barking in the background



- Case-2: Music in the background

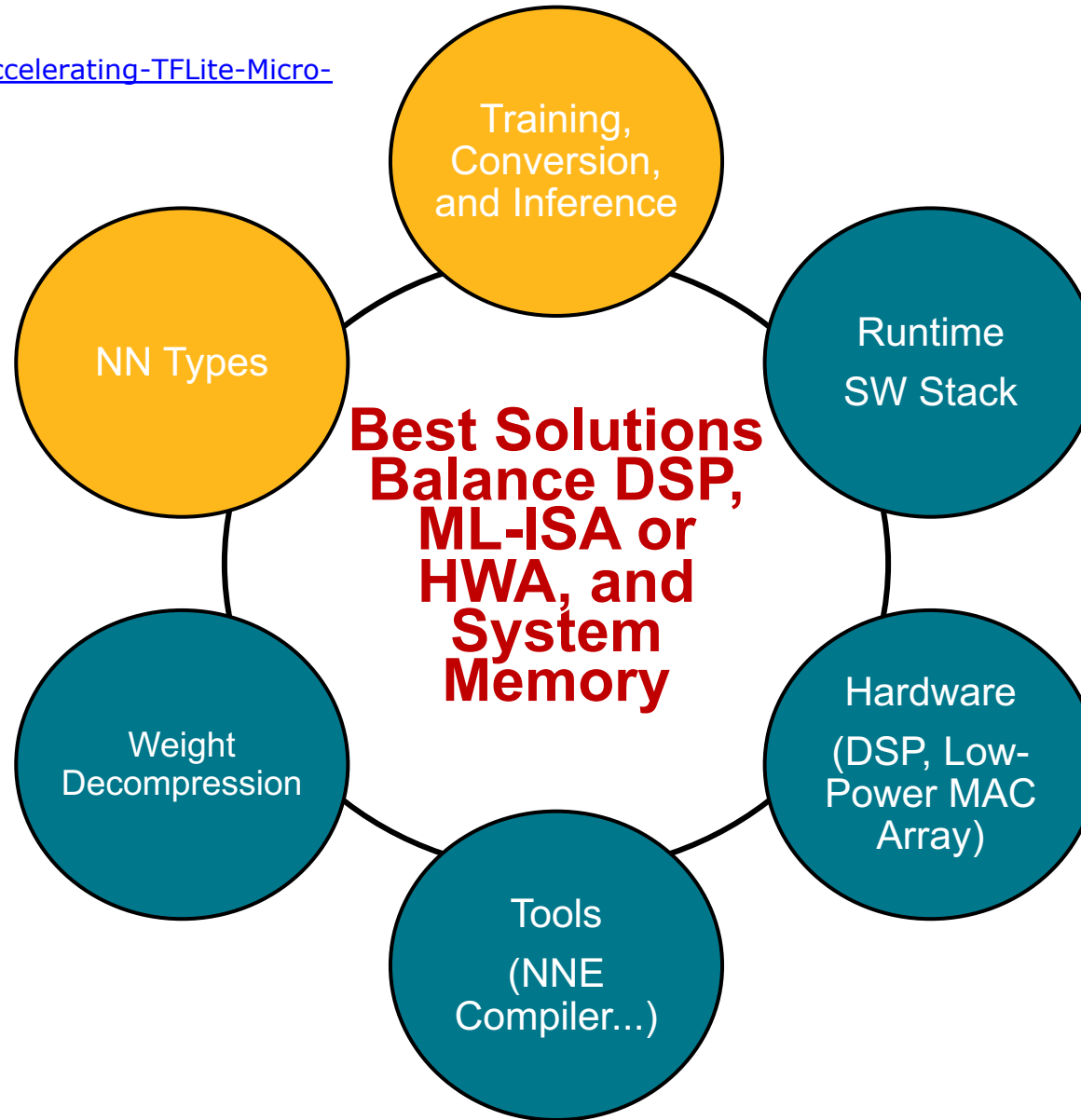


Extending Offload from DSP to *Tiny* NN Engine



Holistic Approach to Realizing a Low-Power Edge AI Platform

<https://blog.tensorflow.org/2022/03/Accelerating-TFLite-Micro-On-Cadence.html>





—

Software Optimization

Creating a Reference C-Based LSTM Operator

A collaborative effort between Google's TFLM and Cadence Audio teams

Basic LSTM operator

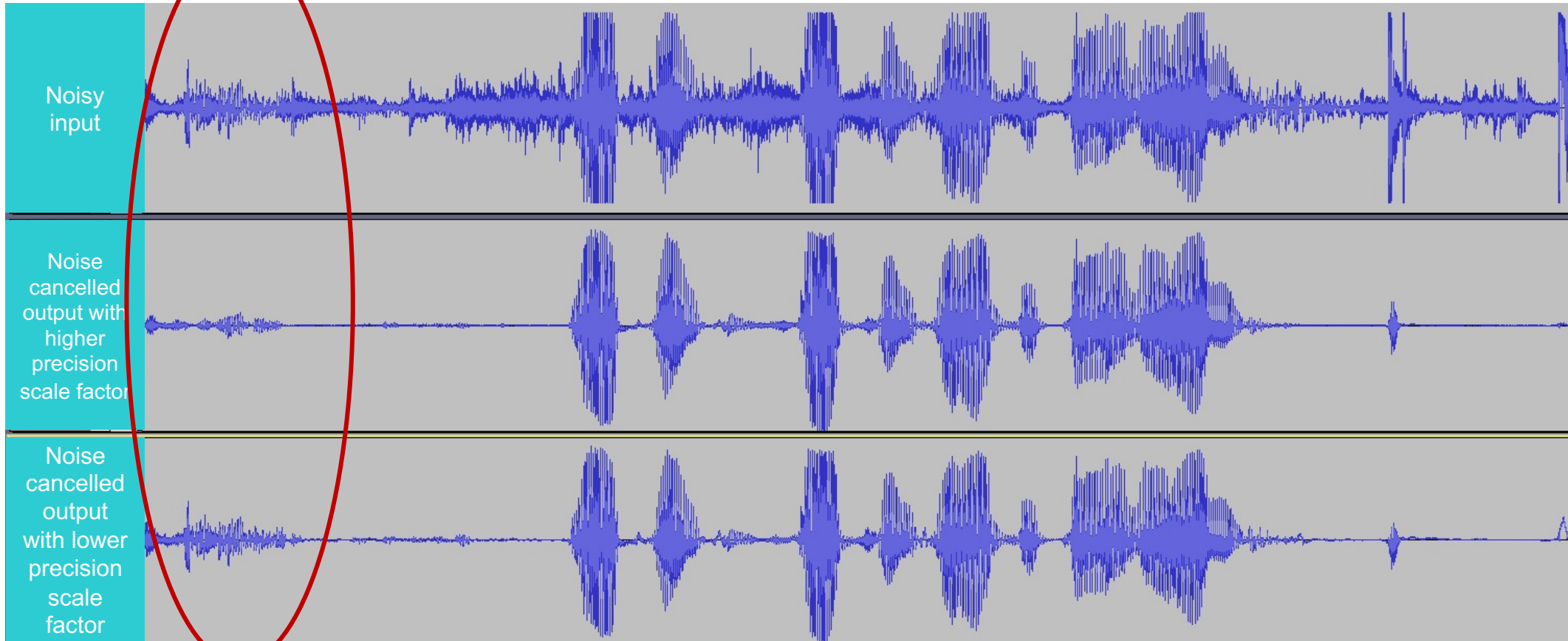
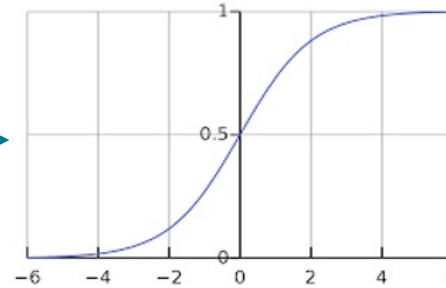
$$\begin{aligned}f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\\tilde{c}_t &= \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \\c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\h_t &= o_t \circ \sigma_h(c_t)\end{aligned}$$

- ✓ There's more to creating Ref C than just implementing in C
- ✓ Smooth flow from training, quantization, to inference
- ✓ Address processor or HWA or both friendly for vector processing (SIMD)
- ✓ Parallel processing

Tensilica® HiFi DSPs first to support LSTM operator in TFLM

How You Allocate Bits Matters: During Float-to-Fixed Conversion

FC (Matrix * Vec)





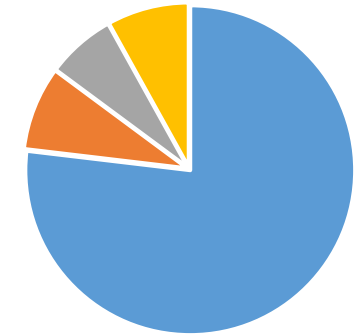
Abstract background featuring a light blue hexagonal grid. In the upper left, a cluster of small white triangles forms a larger hexagonal shape. Scattered across the scene are numerous 3D cubes of various sizes and colors, including red, orange, yellow, and blue, some of which have a bright white highlight. Faint mathematical formulas are visible on some of the grid's hexagonal cells, including $F(s) = -e^{-sT} \left(\frac{\alpha}{s+\alpha} \right)$, $f_m(t) = 1 - e^{-\beta t}$, $F(s) = -e^{-s}$, and $r = m_0 \sqrt{2m_1 - m_0^2}$.

Hardware Optimization

Dissecting a Sample LSTM Operator in Terms of Compute Cycles

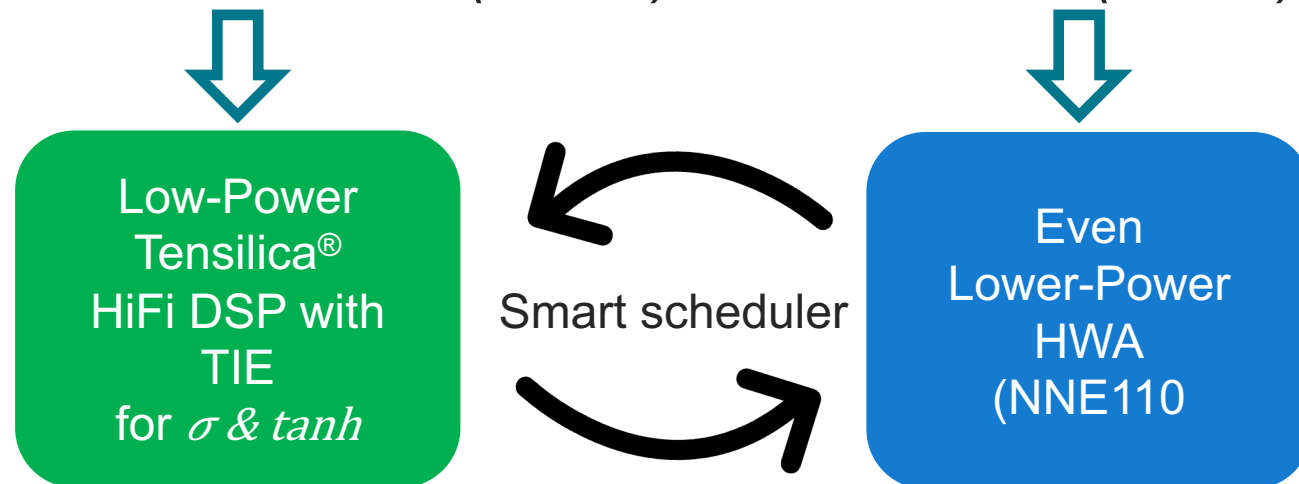
Cycles	Calls per layer	Dimension	% cycle contribution
Mat*Vec (FC)	8	128x128	76.88
Sigmoid	3	128-point	8.26
Tanh	2	128-point	6.8
Elementwise, control code	4	NA	8.06

% cycle contribution
(without TIE)



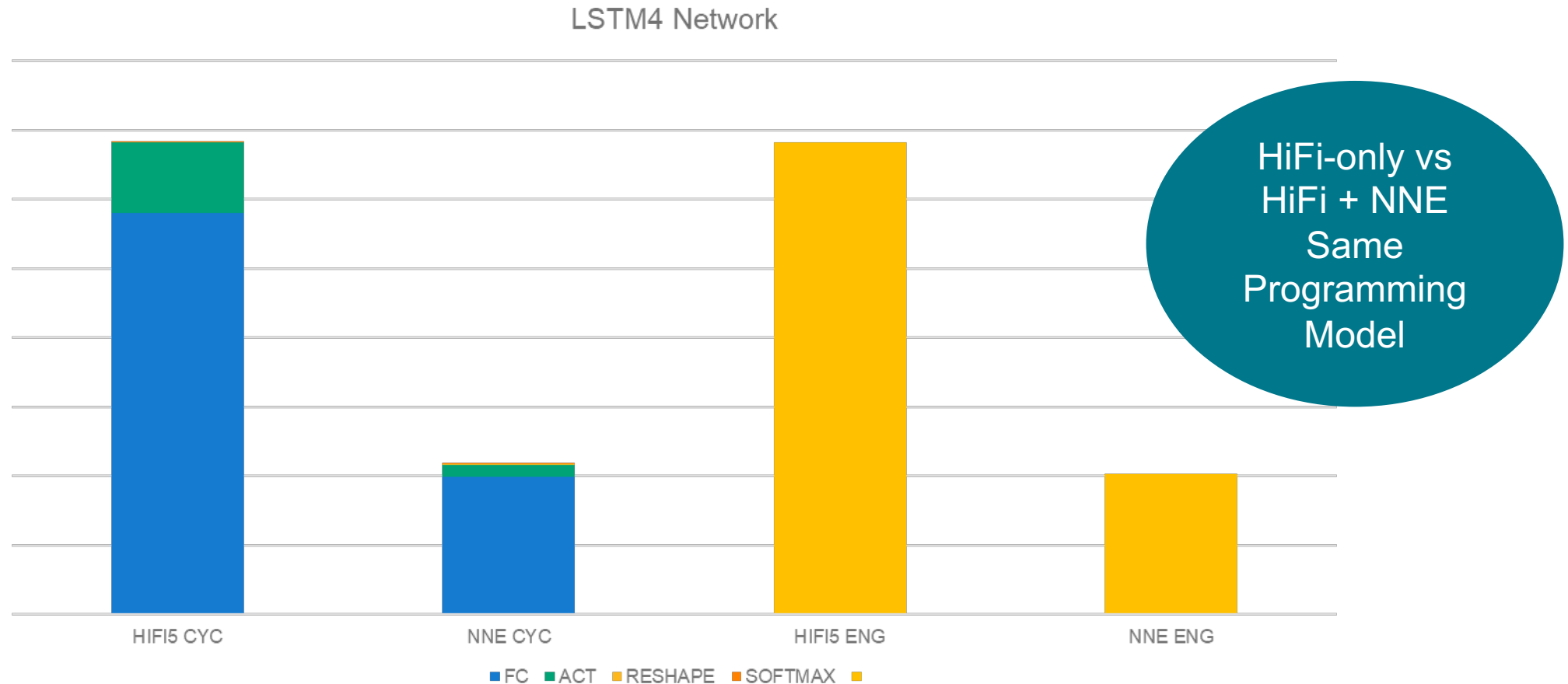
■ Mat*Vec (FC) ■ Sigmoid ■ Tanh ■ Remaining

*σ & tanh functions (~15%) and Mat*Vec (~77%) contribute to ~90% of cycles*



Solving Challenges 1 and 2: Latency and Power

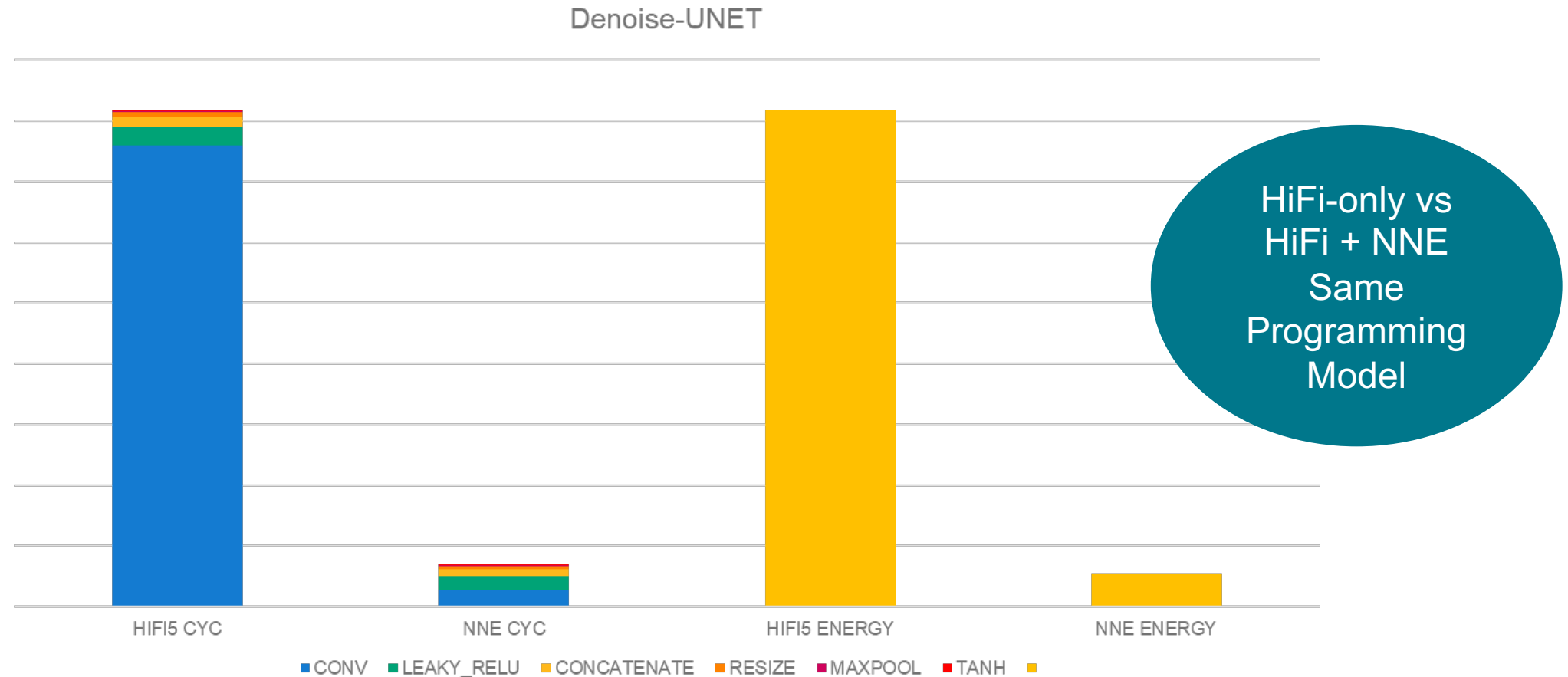
Latency (cycles) reduced by factor of **~3.14X**, while energy reduced by **~3.36X**!



Performance and Energy Measurements for a Sample LSTM NN

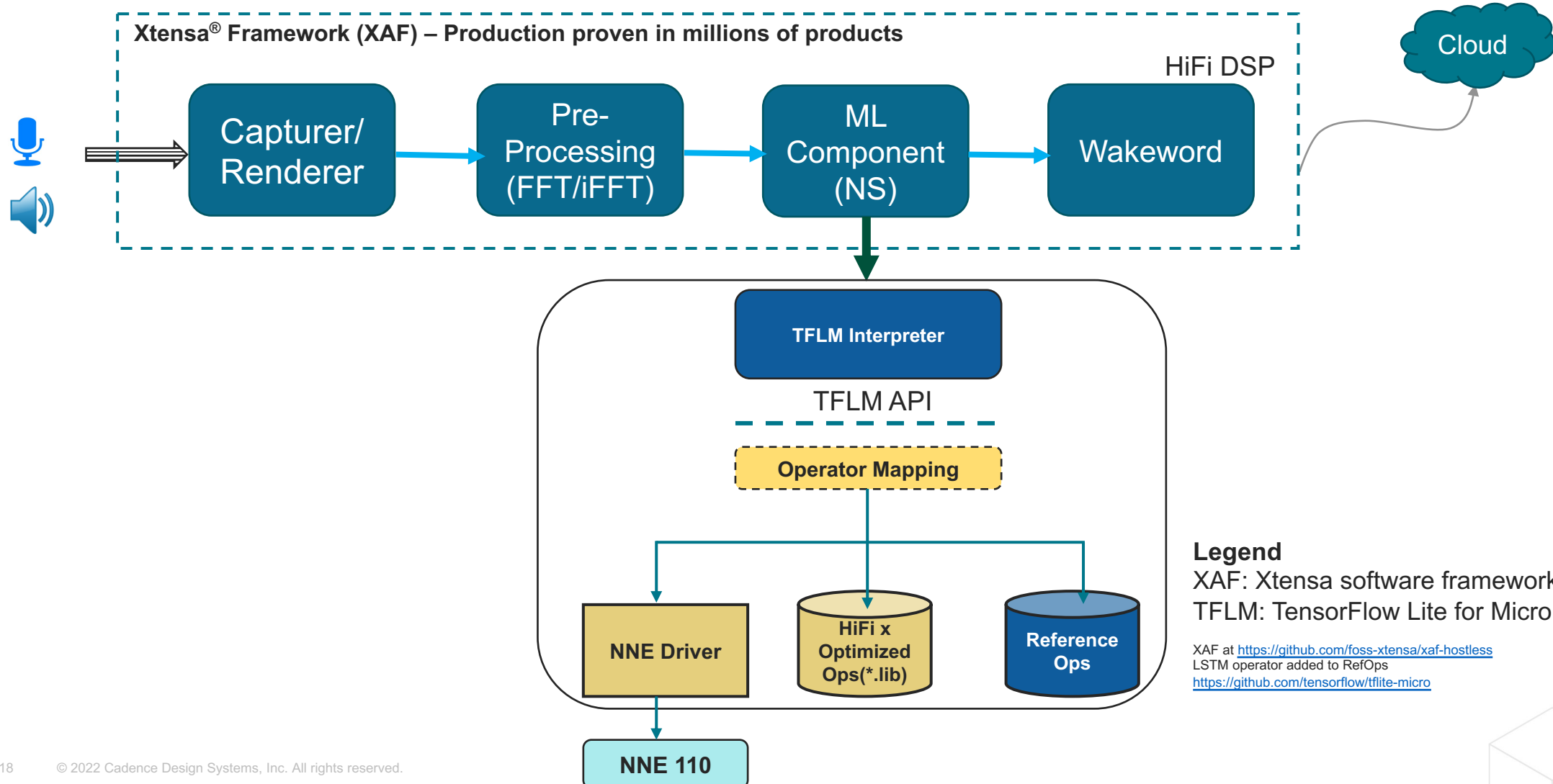
Solving Challenges 1 and 2: Latency and Power

Latency (cycles) reduced by factor of **~12X**, while energy reduced by **~15X!**

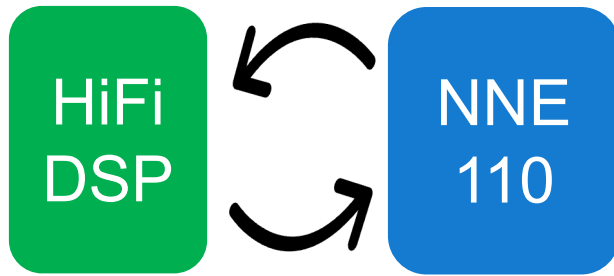


Performance and Energy Measurements for a CNN-Based NS NN

Solving Challenge 3: Combining Audio and ML Algos (XAF + TFLM)



HiFi DSP + NNE110 Achieves Holistic Balance



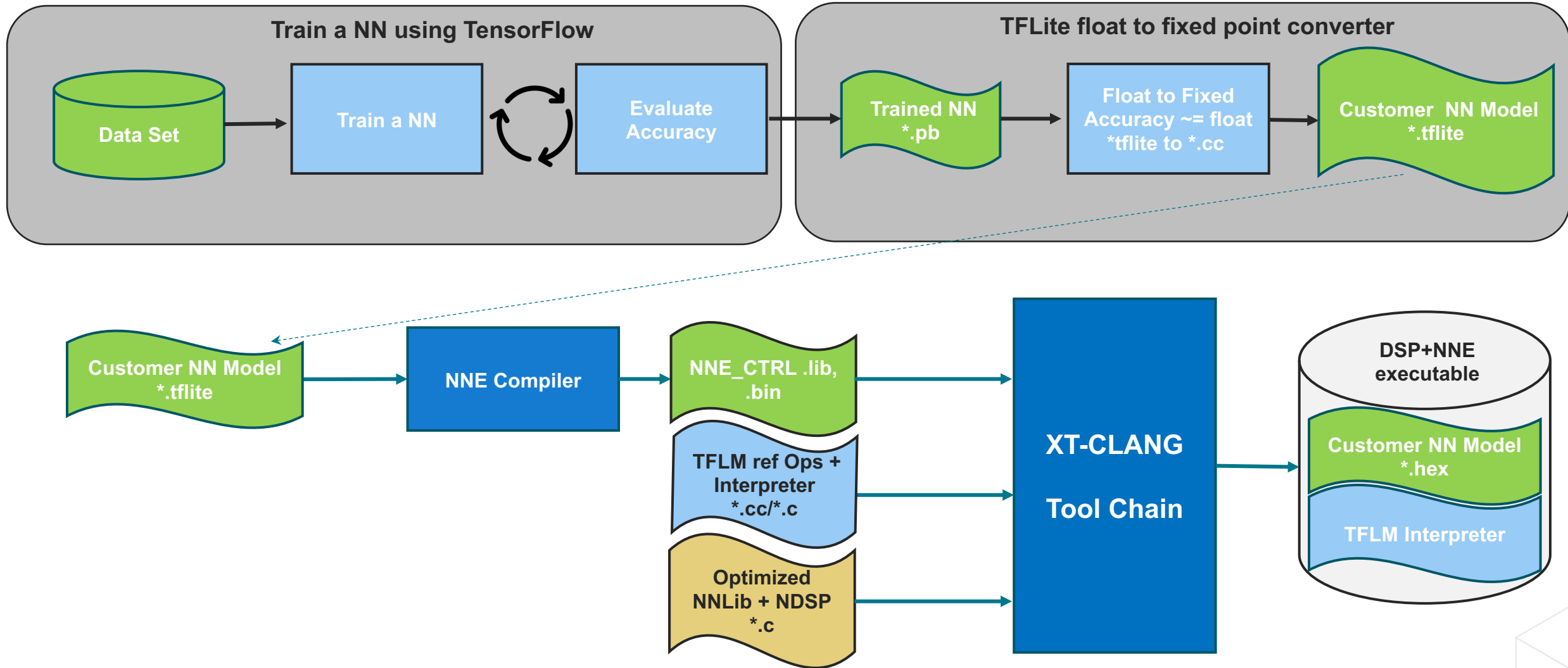
- **NN Types:** CNN, DS-CNN, LSTM
https://github.com/tensorflow/tflite-micro/blob/main/tensorflow/lite/micro/kernels/xtensa/lstm_eval.cc
- **Training Framework:** LSTM
https://github.com/tensorflow/tflite-micro/tree/main/third_party/xtensa/examples/micro_speech_lstm/train
- **Runtime SW stack:**
<https://github.com/tensorflow/tflite-micro>
<https://github.com/foss-xtensa/xaf-hostless>
- **Hardware:** Best of DSP and accelerator
- **Tools:** NNE Compiler, energy-aware scheduler
- **Weights decompression**



cādence®

© 2022 Cadence Design Systems, Inc. All rights reserved worldwide. Cadence, the Cadence logo, and the other Cadence marks found at www.cadence.com/go/trademarks are trademarks or registered trademarks of Cadence Design Systems, Inc. Accellera and SystemC are trademarks of Accellera Systems Initiative Inc. All Arm products are registered trademarks or trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All MIPI specifications are registered trademarks or service marks owned by MIPI Alliance. All PCI-SIG specifications are registered trademarks or trademarks of PCI-SIG. All other trademarks are the property of their respective owners.

Offline Steps to Create a Fixed-Point NN Executable for Inference





AONdevices

arm

ASPINITY

brainchip
The Neuromorphic Computing Company

CEVA®

Deeplite

EDGE IMPULSE

emza
visual sense

FotaHub

GREENWAVES
TECHNOLOGIES

Grovetly Inc.

Himax

HOTC

imagimob

infineon

itemis

KLIKA·TECH
GLOBAL IOT SOLUTIONS

LatentAI

LATTICE
SEMICONDUCTOR

Micro.ai

OmniML

NXP

POI

Plumerai

PROPHESSEE

Qeexo

Qualcomm

Rackner

RealityAI®
Engineering Solutions for the Edge

REXEN
technology

RENESAS

SAP

seeed
The IoT Hardware Enabler

SensiML

Sony Semiconductor
Solutions
Corporation

ST
life.augmented

SA STREAM ANALYZE

synaptics®

SynSense

SYNTIANT

Tensil.ai

TensorFlow

XMOS

Copyright Notice

This presentation in this publication was presented as a tinyML[®] Summit 2022. 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