# Benchmarking AI compiler for the TinyML market

Peter Chang

TinyML Asia 2023, 2023/11/16

# Outline

1. Introduction to the MLPerf Tiny benchmark
2. Optimization of AI models from compiler's perspectives
3. Possibility of the future benchmark designs for TinyML market
4. Epilogue




skymizer

# 1. Introduction to the MLPerf Tiny benchmark

# What is MLPerf benchmark?

The foundation for MLCommons® started with the MLPerf™ benchmarks in 2018, which established industry-standard metrics to measure machine learning performance and…

https://mlcommons.org/en/history/




skymizer

# What is MLPerf Tiny benchmark?

- MLPerf Tiny serves as a Machine Learning Inference benchmark collection tailored for TinyML systems.
- MLPerf enables the assessment of energy consumption and inference speed for AI models focused on visual and audio tasks.
- All submitters must fit the quality targets for each use case for close division.

| Abbr. | Use Case | Model | Quality Target | Parameters |
|---|---|---|---|---|
| **AD** | Anomaly Detection | Deep AutoEncoder | 0.85 (AUC) | 270 kPar |
| **KWS** | Keyword Spotting | DS-CNN | 90% (Top 1) | 52 kPar |
| **IC** | Image Classification | ResNet | 85% (Top 1) | 96 kPar |
| **VWW** | Visual Wake Words | MobileNet | 80% (Top 1) | 325 kPar |

| Typical Systems | |
|---|---|
| Processor | **MCU (+ DSP/NPU)** |
| Frequency | **10s-100s MHz** |
| Memory | **MB Flash & SRAM** |
| Power | **mW Power** |
| AI Model Size | **< 1M Parameters** |

https://www.youtube.com/watch?app=desktop&v=i4wCqoVcdJI




skymizer

# Submission Requirements

- Submitter
- Device Description
- Processor and count
- Accelerator and count
- Software
- Benchmark Results
- Code
- Accuracy/AUC

**Notice:**

Energy Number is optional.

**System Category:**

1. Available System
2. Preview Systems
3. Research, Development, or Internal (RDI)

Available System Category comprise solely of components that can be obtained for purchase or leased from cloud services.



skymizer

# Submission Process

## Submission

- Sign CLA
- Provide POCs with Github handles and email addresses

## Review

- All submitters are peer-reviewers
- Reviewers fill objection opinions
- Peer review objections
- Submitters revise based on objections
- Vote for accept or not

## Publication

- Write Supplemental materials to describe your work
- Join press conference meeting before publish
- Release the results on the MLCommons website

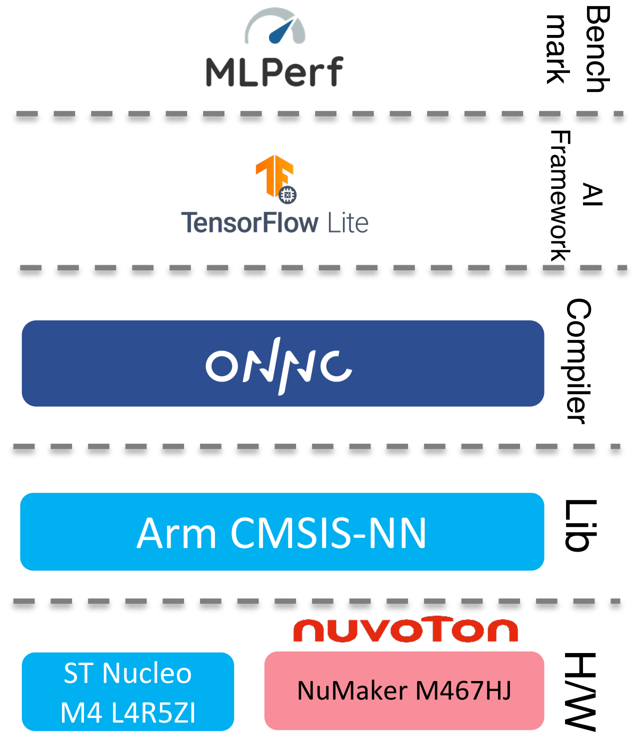


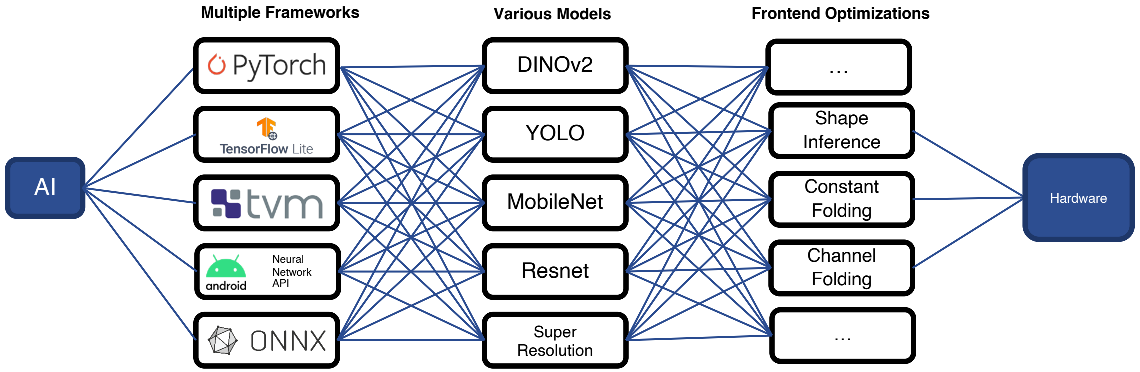

skymizer

# 2. Optimization of AI models from compiler's perspectives

# ONNC & Our Submissions

- ONNC is an AI compilation suite developed by Skymizer for various markets, from cloud to tiny devices.
- For tiny devices, ONNC compiles AI models to C codes which call Neural Network Library for the target board.
- We use ONNC to compile AI models to C codes for Nuvoton NuMaker M467HJ Cortex-M4 board and the benchmark's reference board.



# MLPerf Submissions on MLPerf v1.1

*Skymizer submit two numbers, one for Skymizer and another as the agent for Nuvoton which is Skymizer's collaboration partner.*

## MLPerf Inference – Tiny

*"The outstanding results achieved in the MLPerf Tiny Benchmark's Cortex-M4 MCU segment highlight Nuvoton's and Skymizer's dedication to pushing the boundaries of machine learning performance in resource-constrained environments."*

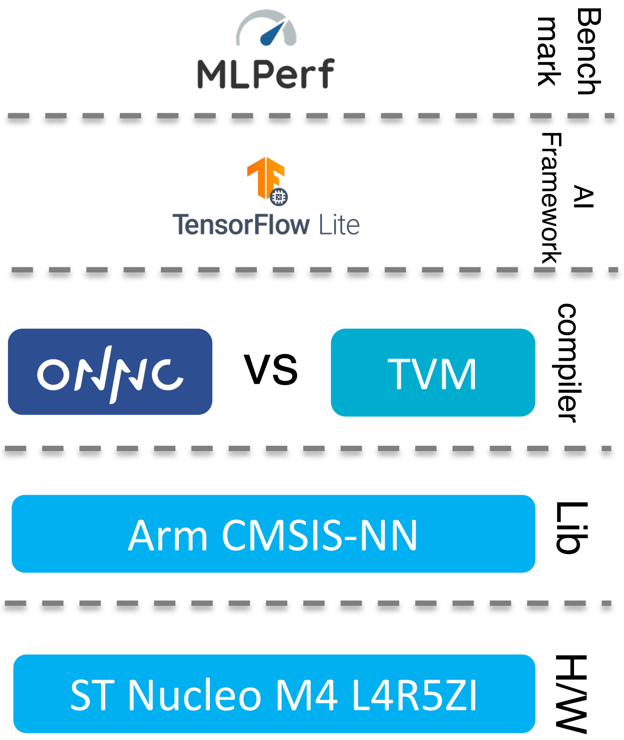
https://www.nuvoton.com/news/news/all/TSNuvotonNews-000456

MLPerf™ Tiny v1.1 Results : closed

| ID | Submitter | System Desc | Board Name | Software |
|---|---|---|---|---|
| **Available** | | | | |
| 1.1-0001 | Krai | NUCLEO-H7A3ZI-Q-X-CUBE-AI-7.3 | NUCLEO-H7A3ZI-Q | X-CUBE-AI v7.3.0 |
| | | NUCLEO-H7A3ZI-Q-X-CUBE-AI-8.0 | NUCLEO-H7A3ZI-Q | X-CUBE-AI v8.0.0 |
| 1.1-0002 | Krai | NUCLEO-L4R5ZI-MICROTVM-CMSIS-NN | NUCLEO-L4R5ZI | microTVM |
| | | NUCLEO-L4R5ZI-MICROTVM-NATIVE | NUCLEO-L4R5ZI | microTVM |
| | | NUCLEO-L4R5ZI-X-CUBE-AI-7.3 | NUCLEO-L4R5ZI | X-CUBE-AI v7.3.0 |
| | | NUCLEO-L4R5ZI-X-CUBE-AI-8.0 | NUCLEO-L4R5ZI | X-CUBE-AI v8.0.0 |
| 1.1-0003 | Krai | NRF5340-DK-MICROTVM-CMSIS-NN | nRF5340 DK | microTVM |
| 1.1-0004 | Nuvoton | NUMAKER-M467HJ-zephyr | NUMAKER-M467HJ | ONNC |
| 1.1-0005 | STMicroelectronics | NUCLEO-H7A3ZI-Q | NUCLEO-H7A3ZI-Q | X-CUBE-AI v8.1.0 |
| 1.1-0006 | STMicroelectronics | NUCLEO-L4R5ZI | NUCLEO-L4R5ZI | X-CUBE-AI v8.1.0 |
| 1.1-0007 | STMicroelectronics | NUCLEO-U575ZI-Q | NUCLEO-U575ZI-Q | X-CUBE-AI v8.1.0 |
| 1.1-0008 | Skymizer | NUCLEO-L4R5ZI-mbed-os | NUCLEO-L4R5ZI | ONNC |
| | | NUCLEO-L4R5ZI-zephyr | NUCLEO-L4R5ZI | ONNC |

https://mlcommons.org/en/inference-tiny-11/

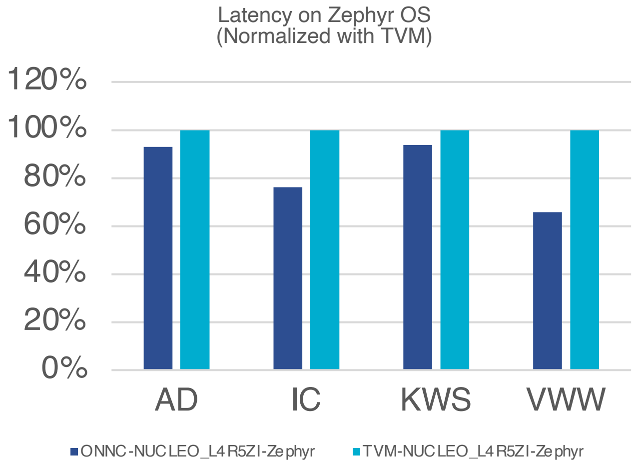
# ONNC Optimization in Latency and Energy




## Software Stack Comparison between ONNC and TVM

## Reduce Latency

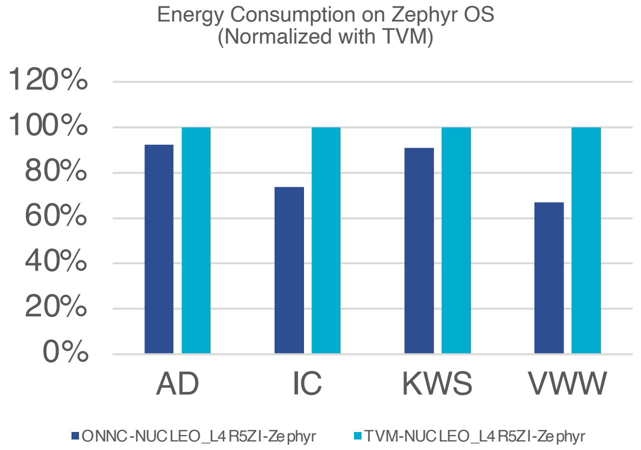
Our latency is **35% less** in the best-case scenario.




| Latency (ms/inf) | AD | IC | KWS | VWW |
|---|---|---|---|---|
| ONNC | 8 | 296.6 | 93.6 | 197.8 |
| TVM | 8.6 | 389.5 | 99.8 | 301.2 |

## Reduce Energy

Our energy is **32% less** in the best-case scenario.




| Energy uJ/inf. | AD | IC | KWS | VWW |
|---|---|---|---|---|
| ONNC | 409.666 | 14927.33 | 4747.946 | 10412.796 |
| TVM | 443.2 | 20236.3 | 5230.3 | 15531.4 |

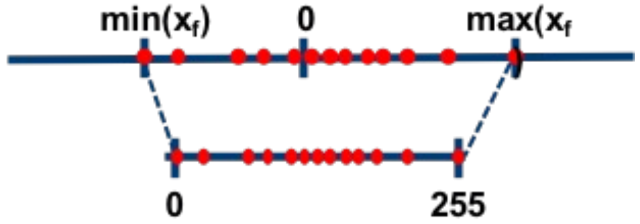
| Abbr. | Use Case | Model | Quality Target |
|---|---|---|---|
| **AD** | Anomaly Detection | Deep AutoEncoder | 0.85 (AUC) |
| **IC** | Image Classification | ResNet | 85% (Top 1) |
| **KWS** | Keyword Spotting | DS-CNN | 90% (Top 1) |
| **VWW** | Visual Wake Words | MobileNet | 80% (Top 1) |




skymizer

# Key components in an AI model compiler

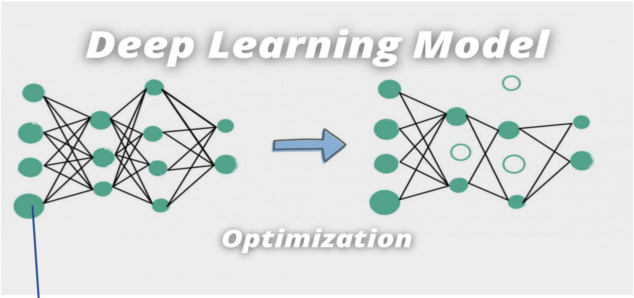
## Quantization




https://intellabs.github.io/distiller/algo_quantization.html

- Convert higher-bit computation into lower-bit computation.

## Graph-Level Optimization




https://www.thinkautonomous.ai/blog/deep-learning-optimization/

- Hardware-friendly.
- Speed up.

## Operator-Level Optimization

```
for(x=0;x<F;x++){
 for(y=0;y<E;y++){
  for(k=0;k<C;k++){
    for(i=0;i<R;i++){
      for(j=0;j<S;j++){
        for(m=0;m<M;m++){
          Output[m][x][y] +=
              Input[k][x+i][y+j] ×
              Weight[m][k][i][j]
} } } } } }
```

- Speed up.
- Hardware instructions.

https://www.intechopen.com/chapters/60223



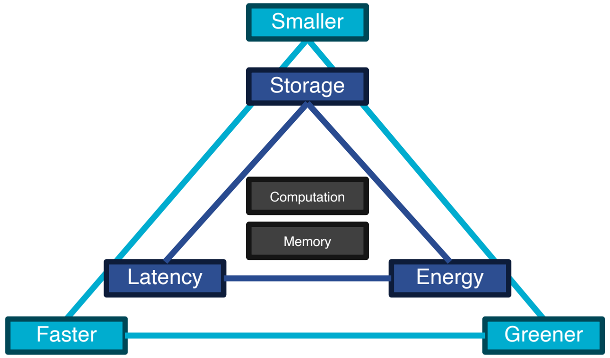
skymizer

# 3. Possibility of the future benchmark designs for TinyML market

## More Energy-Centric

For tiny device developers, energy consumption usually will be the first key factor to decide whether they should adapt AI or not.

Designing a more energy-centric/energy-priority benchmark would fit developers' needs more.




p52, Lecture 2, "TinyML and Efficient Deep Learning Computing", S. Han, 2023

## More Whole-System's View

The whole system benchmark can show the performance and energy numbers not only from AI inferencing but also from pre-/post-processing and OS.

The whole system performance and energy analysis will also faciliate their decisions for those who try to decide which evaluation board to buy.

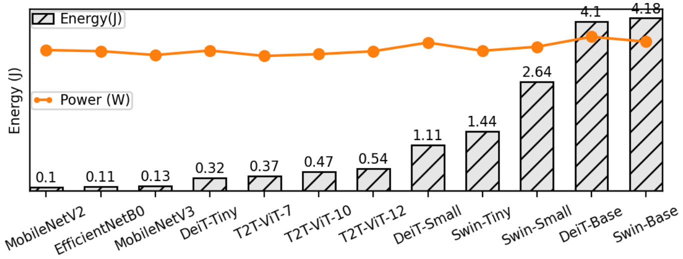


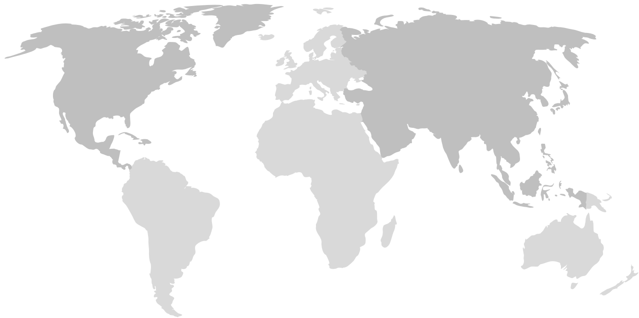

**Figure 5: Energy and power consumption on Pixel 4.**

Wang, Xudong, et al. "Towards efficient vision transformer inference: A first study of transformers on mobile devices." *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications.* 2022.

## More Comprehensive

AI benchmarks for tiny devices would be better if it can cover not only audio and visual but also other more sensing modalities, like environmental sensing modalities.

Also, for data sets, an open and comprehensive data set suite would be more realistic, more comprehensive and be more non-discrimination.






skymizer

# 4. Epilogue

## ONNC RISC-V Support & Booth

- ONNC also support RISC-V as MCU.
- We have a demo using Tinker V board with Andes IP in the booth area.

## MLPerf Tiny Next Round Submission

- Submission: February 23th, 2024 (Planning)
- Publication: March 27th, 2024 (Planning)
- Would add 2 streaming benchmarks (Planning)
  - Streaming Denosing LSTM & Streaming KWS

## LLM on tiny devices

- Model compression with accuracy ensurance will be the key in landing LLM or Transformer-based models to tiny devices.
- Skymizer also have set this agenda within our roadmap.



skymizer

TINY ML ASIA
Seoul, South Korea
November 16, 2023

# Copyright Notice