Powering innovation in a new world of AI devices on the edge with microNPUs

Tanuj Arora
Product Manager, Ethos-U microNPUs’
Machine Learning Group
AI is transforming... everything

- Security cameras
- Retail
- Smart home, smart cities
- Public transportation
- Consumer home electronics
- Home hub
Extending the Arm Endpoint AI solution

Providing NN acceleration in highly constrained environments

Ubiquitous presence  
NN acceleration in software

Orders of magnitude increase in NN perf  
Easy integration into existing design

Signal Processing  
Neural network acceleration

Common software development environment secures any investment made on software development
Ethos-U55: First microNPU for Cortex-M CPUs

- Neural network processor for Cortex-M systems
  - Works alongside Cortex-M55, Cortex-M7, Cortex-M33 and Cortex-M4 processors
- Designed for embedded type systems
  - Fast on-chip SRAM and a slower system flash
- Heavy compute operators for CNN and RNN accelerated in hardware.
- Support for efficient weight compression
  - Compression typically offline
  - Decompression on-the-fly
- Support for 8-bit or 16-bit activations
  - Weights are always 8-bit
- Configurations 32, 64, 128 or 256 MAC/cc
  - 8-bit activations use 1 cc per MAC
  - 16-bit activations use 2 cc per MAC
Ethos-U65: Powering innovation in a new world of AI devices

- 512 MACs/cc and 256 MACs/cc configurations
- Designed to run in Cortex or Neoverse based systems and is tolerant to high DRAM latencies
- Dual AXI
  - SRAM (128-bit AXI) and DRAM (128-bit AXI) support. A-class memory system support
  - Better for weight bound networks
- Able to run any size networks with a combination of SRAM and DRAM
  - Dram spilling splits activations between DRAM and SRAM.
Different systems applicable to Ethos-U65
# High level differences between Ethos-U55 and Ethos-U65

Both are instantiations of the same architecture

<table>
<thead>
<tr>
<th>Ethos-U55</th>
<th>Ethos-U65</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 configs: 32/64/128/256 MACs/cycle</td>
<td>2 configs: 256/512 MACs/cycle</td>
</tr>
<tr>
<td>Designed for SRAM + flash</td>
<td>Designed for SRAM + DRAM and/or flash</td>
</tr>
<tr>
<td>Host CPU support: Cortex-M55, Cortex-M7, Cortex-M4 and Cortex-M33</td>
<td>Host CPU support: Cortex-M55 and Cortex-M7</td>
</tr>
</tbody>
</table>
| Two 64-bit AXI master interfaces  
M0: Full read+write AXI master to SRAM  
M1: Read only AXI master to flash | Two 128-bit AXI master interfaces  
M0: Full read+write AXI master to SRAM  
M1: Full read+write AXI master to DRAM |
Network support in Ethos-U

• With the supported operator set, Ethos-U can completely execute most of the popular networks today. For ex:

  • For networks that cannot be executed on Ethos-U completely, the operators unsupported by Ethos-U fallback to the attached Cortex-M CPU
    • These are accelerated through CMSIS-NN library

...and many more
Cortex-M optimized software flow

- Train network in TensorFlow
- Quantize it to Int8 TFL flatbuffer file (.tflite file)

- Runtime executable file on device
- The NN is executed on Cortex-M
  - CMSIS-NN optimized kernels if available
  - Fallback on the TFLu reference kernels
Ethos-U optimized software flow

- Train network in TensorFlow
- Quantize it to Int8 TFL flatbuffer file (.tflite file)
- NN Optimizer identifies graphs to run on Ethos-U
  - Optimizes, schedules and allocates these graphs
  - Lossless compression, reducing size of tflite file
- Runtime executable file on device
- Accelerates kernels on Ethos-U. Driver handles the communication
- The remaining layers are executed on Cortex-M
  - CMSIS-NN optimized kernels if available
  - Fallback on the TFLu reference kernels

TFL: TensorFlow Lite
Ethos-U65 software flow (A-class based system)

- TFLµ-based flow same as Ethos-U55
  - Example Linux compatible A-class inference
  - Example API A-class software subsystem driver
  - Example M-class wrapper app
  - Ethos-U65 driver
- Optimization strategies for DRAM and SRAM
Performance Results: Ethos-U55

- Ethos-U55 provides an exponential increase over previous generation Cortex-M
- 128 MACs/cycle config used for these calculations. Bigger configuration of 256 MACs/cycle is also available.

- DS-CNN-L: Keyword Spotting
- MobilNet: Image Classification
- Wav2letter: Speech Recognition
- Model is stored in Flash.
- 4GB/s SRAM Bandwidth | 500MB/s Flash Bandwidth
- Host CPU for Ethos-U55 is Cortex-M55
- CMSIS_NN Optimized kernels used for Cortex-M55 performance
Ethos-U65 performance improvements

2.5x Average network performance improvements (Infr/S) with increased MAC/Cycle and IP Memory bandwidth

Ethos-U65 relative performance improvement over Ethos-U55

Higher bandwidth = far better performance

- Voice Key Word Spotting (DS-CNN-L)
- Vision Feature Extraction (Mobilenet v2 1.0 224)
- Vision Feature Extraction (Mobilenet v1 1.0 224)
- Human Activity Recognition (HAR CNN)
- Audio Enhancement (RNNoise)
- Automatic Speech Recognition (wav2letter)

Ethos-U65:
- SRAM bandwidth 8 GB/s
- DRAM Bandwidth 1.875 GB/s

Ethos-U55:
- SRAM bandwidth 4 GB/s
- Flash Bandwidth 500 MB/s
Cortex-A53 vs Ethos-U65 (MobileNet v2 1.0 224)

With Ethos-U65, we get

11.9x times performance

89.8x times power efficiency

Over Cortex-A53
Corstone-300 Ecosystem FVP for Least Investment Path to SW

- **Available on Arm Developer** at no cost
  - Cortex-M55 Corstone-300 FVP (MPS2 based)
    - Blog: Introduction to Corstone-300 Ecosystem FVP
  - Ethos-U55 Corstone-300 FVP (MPS3 based)

- Ethos-U software
  [https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ethos-u/](https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ethos-u/)

Peripherals are representative, please refer to documentation of the specific FVP for its peripherals.
Powering innovation in a new world of AI applications

- Extending performance efficiency
- Flexible integration
- Simple and consistent Cortex-M software development
- Industry-leading ecosystem

Cortex-M today

Cortex-M and Ethos-U55 or Ethos-U65

Data throughput

Vibration detection | Sensor fusion | Keyword detection | Anomaly detection | Object detection | Gesture detection | Biometric awareness | Speech recognition | Object classification | Real-time recognition

TOP/s
Thank You
Danke
Merci
Merci
Merci
あ Hawaiia
Gracias
Kiitos
감사합니다
ধন্যবাদ
شكرًا
ধন্যবাদ
תודה
Sponsors

Premier Sponsor & tinyML Strategic Partner

arm

Gold Sponsor

EDGE Cortex

Silver Sponsor

SynSense
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

Profiling and debugging tooling such as Arm Keil MDK

Application

Optimized models for embedded

Runtime (e.g. TensorFlow Lite Micro)

Optimized low-level NN libraries (i.e. CMSIS-NN)

RTOS such as Mbed OS

Arm Cortex-M CPUs and microNPUs

AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
Dynamic Neural Accelerator™

- 10x more compute with single DNA engine
- More than 20x better energy-efficiency
- Ultra-low latency
- Fully-programmable with INT 8bit support

www.edgecortix.com

TARGET MARKETS
- Automotive
- Robotics
- Drones
- Smart Cities
- Industry 4.0

Tight coupling between software & hardware with automated co-design

© 2020 EDGECORTEX. ALL RIGHTS RESERVED
SynSense builds **ultra-low-power** (sub-mW) **sensing and inference** hardware for **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
Partners

Conference Partner

Media Partners

[Logos of various partners and media partners]
Questions?
Or to join tinyML WeChat Group

Please add staff to join our official tinyML WeChat Group
Copyright Notice

This presentation in this publication was presented at tinyML® Asia 2020. 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