MediaTek NeuroPilot - Micro

Rituparna Mandal and Max Wu – MediaTek

India Area Group – April 22, 2021
tinyML Talks Sponsors

arm
tinyML Strategic Partner

EDGE IMPULSE

maxim integrated™

Qeexo

Reality AI®

SynSense

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

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

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

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

© 2020 Arm Limited (or its affiliates)
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
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

- Embedded and edge compute deployment options

**Impulse**
- Test impulse with real-time device data flows

**Test**
- Real sensors in real time
- Open source SDK
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](http://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](http://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](http://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
Add Advanced Sensing to your Product with Edge AI / TinyML

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

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
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
Successful tinyML Summit 2021:

- **5** days of tutorials, talks, panels, breakouts, symposium
  - 4 tutorials
  - 6 keynotes & 6 plenary tinyTalks (more in breakouts)
  - 2 panel discussions
  - 5 disruptive news presentations
  - 17 breakout/partner sessions
  - 6 Best Product and Innovation Award Finalists & Presentations
  - 89 Speakers

- **5006** registered attendees representing:
  - 104 countries, 1000+ companies and 400+ academic institutions

- **26** Sponsoring companies

www.youtube.com/tinyML with 150+ videos

tinyML Summit-2022, January 24-26, Silicon Valley, CA
June 7-10, 2021 (virtual, but LIVE)
Deadline for abstracts: May 1

Sponsorships are being accepted: sponsorships@tinyML.org
## Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, April 27</td>
<td>Michael Jo and Xingheng Lin</td>
<td>Train-by-weight (TBW): Accelerated Deep Learning by Data Dimensionality Reduction</td>
</tr>
<tr>
<td></td>
<td>Rose-Hulman Institute of Technology</td>
<td></td>
</tr>
</tbody>
</table>

Webcast start time is 8 am Pacific time

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

Chetan Singh Thakur, PhD
Assistant Professor at the Indian Institute of Science (IISc), Bangalore. He is a Ph.D. in neuromorphic engineering. Dr. Thakur's research interest spans VLSI Design, Edge Computing, Neuromorphic Engineering.

Abhishek Nair
Abhishek is a PhD student at IISc Neuronics lab. His research area includes exploring low power ML algorithms for digital hardware implementation. He has 5+ years of semiconductor industry experience.

Sandipan Chatterjee
Sandipan is Lead Data scientist at DXC Technology where he develops and implements vision-based automation in manufacturing, automotive and healthcare. He has a background in image and statistical analysis.

Anup Rajput
Co-founder at Envir AI, trying to bring ML into the real world. Anup has a background in semiconductor design and applied ML from edge to cloud.

Follow us for more updates at: https://www.linkedin.com/company/tinyml-india
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Rituparna Mandal

• Rituparna Mandal is the General Manager of MediaTek Bangalore. In addition, she is the Director of the Advanced Technology team where she is responsible for foundation IP and advanced CPU designs across MediaTek products. Rituparna is a B. Tech. in Electronics Engineering from Jadavpur University. She has over two decades of experience in the semiconductor industry.
Max Wu

Max Wu is the Director of the Computing and Artificial Intelligence Technology Group at MediaTek. He is currently responsible for AI technology and strategy planning. He is leading the team to explore AI applications in mobile, home entertainment, smart surveillance, health care, and new business opportunity. Max has more than 15 years of experience in semiconductor SoC development. He received his M.S. degree from the Institute of Electrical and Control Engineering, National Chiao-Tung University, Taiwan.
tinyML Talk – MediaTek NeuroPilot-Micro

Rituparna Mandal & Max Wu

22-Apr-2021
Agenda

- MediaTek Brief
- Edge AI and Challenges
- MediaTek Solution
- Our Vision
Balanced Product Portfolio

cross-platform, global market leadership

revenue ratio*:

- **Mobile Computing**: 45-50%
- **Growth**: 29-33%
- **Smart Home Related**: 21-26%

### Mobile Computing

- Android Tablet
- Smartphone
- Arm-based Chromebook
- Voice Assistant Device

### Growth

- Connectivity, Broadband & Networking
- Power Management
- ASIC

### Smart Home Related

- Smart TV & Digital TV
- Optical Drive & BD Player
- Feature Phone

#1 #1 #1 #1 growing market share #1 #1 #1

Source (ranking by 2020 market share): Strategy Analytics, Gartner, IDC, IC Insight, iSuppli and MediaTek company data.

* based on 2020 Q4 revenue ratio
Aiming at Future

shift in scale reflecting the potential opportunity
Cloud and Edge Collaboration Will Be the Trend

- Cloud centric approach is limited by network traffic, latency and privacy concern.
- Computing at the edge improves significantly driven by mobile phone technology in the past decade
- 5G and edge AI are the key enabling technologies to our intelligent future
Edge AI Device - Design Challenges

- Emerging APPs demand for higher computation complexity and communication traffic
- Need concurrent incorporation of multiple processing engines (modem, APU, CPU, GPU)
- Processing engines generate heat and consume DRAM bandwidth
### NeuroPilot – MediaTek Edge AI Platform

<table>
<thead>
<tr>
<th>High Integration SoC with AI Processing Engine</th>
<th>Best System Performance and Power Efficiency</th>
<th>Flexibility and Easy to Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>• CPU/GPU/APU/DSP</td>
<td>• Heterogeneous computing</td>
<td>• Support main stream AI frameworks (TF, Caffe, ONNX)</td>
</tr>
<tr>
<td>• ISP, Codec, Connectivity</td>
<td>• SDK and toolkits for platform aware optimization</td>
<td>• Cross OS (Android/Linux/RTOS)</td>
</tr>
</tbody>
</table>

#### Framework (TensorFlow / Caffe / Caffe2 / ONNX / MXNet....) API

- Security
- Face
- Video
- Interaction
- Voice

- High Integration SoC with AI Processing Engine
- Best System Performance and Power Efficiency
- Flexibility and Easy to Use

- Android, Linux, RTOS

- Support main stream AI frameworks (TF, Caffe, ONNX)
- Cross OS (Android/Linux/RTOS)

#### SDK & toolkits
- CPU
- GPU
- APU
NeuroPilot vs. NeuroPilot-Micro

Training

TensorFlow
NeuroPilot MLKit (Quantization, Pruning)
TensorFlow Lite

Inference

tflite

NeuroPilot

Model file to C array

NeuroPilot-Micro

CPU Opt. Lib
GPU Runtime/Driver
APU Driver

Edge

CPU
GPU
APU

NeuroPilot-Micro

HW Driver
HW accelerator

MCU Opt. lib

Copyright © MediaTek Inc. All rights reserved. 2021-04-23
# NeuroPilot + i500/i350/i300 Platform (Edge AI)

## Heart Rate Detection on i500

<table>
<thead>
<tr>
<th>Solution</th>
<th>AI Performance</th>
<th>Processor Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>i500</strong></td>
<td>0.75 TOPs</td>
<td>4x CA73 (2.0GHz) + 4x CA53 (2.0GHz)</td>
</tr>
<tr>
<td><strong>i350</strong></td>
<td>0.45 TOPs</td>
<td>4x CA53 (2.0GHz)</td>
</tr>
<tr>
<td><strong>i300</strong></td>
<td>0.08 TOPs</td>
<td>4x CA35 (1.5GHz)</td>
</tr>
</tbody>
</table>
## Some spec details

<table>
<thead>
<tr>
<th></th>
<th>i300</th>
<th>i500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td>28nm</td>
<td>12nm</td>
</tr>
<tr>
<td><strong>Apps CPU</strong></td>
<td>4 x Cortex-A35 1.5GHz</td>
<td>4x A73 + 4x A53 (2.0GHz) (2.0GHz)</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>32-bitx1 LPDDR3/DDR3L 1600Mbps DDR4 1333 Mbps</td>
<td>32-bit LPDDR3 16x2ch LPDDR4X 3733Mbps</td>
</tr>
<tr>
<td><strong>Camera</strong></td>
<td>8MP@30fps</td>
<td>3x ISP 32 MP@30fps Tensilica VP6x2 Depth, FE/FM, Motion</td>
</tr>
<tr>
<td><strong>Video Decode</strong></td>
<td>1080p 60fps H.265 / H.264 / VC-1</td>
<td>1080p@30fps H.264/265 4K@30fps H.264/265 5W</td>
</tr>
<tr>
<td><strong>Video Encode</strong></td>
<td>720p 30fps H.264</td>
<td>1080p 30fps H.264</td>
</tr>
<tr>
<td><strong>Graphics</strong></td>
<td>IMG GE8300 500MHz</td>
<td>MALI-G72 MP3 700MHz</td>
</tr>
<tr>
<td><strong>Display</strong></td>
<td>1920x1200 60fps MIPI / LVDS / HDMI</td>
<td>FHD 1920x1200 60fps</td>
</tr>
<tr>
<td><strong>Connectivity</strong></td>
<td>Integrated Wi-Fi bgn/ BT MT7668(Optional) WiFi ac/abgn/BT5.0</td>
<td>Integrated 1x1 AC/b/g/n GPS/ BT4.2 MT7668(Optional) WiFi 2x2 ac/abgn/BT5.0</td>
</tr>
</tbody>
</table>
NeuroPilot-Micro + MT3620 Platform (tinyML)

Shelf Status Detection
NeuroPilot-Micro – Dual-Core Processing

160x120 RGB image

Results display

Object detection

Concurrently for real-time

Person detection

Tiny vision AI @ Dual 200MHz CM4F
Performance Benchmark

- MTK IP provides good performance for tinyML
- @ MTK SoC, the power of model running on DRAM with NeuroPilot-Micro has similar power as TCM
ML for Systems and Systems for ML

Systems

NeuroPilot-Micro
NeuroPilot

Machine Learning

Systems

NeuroPilot

Framework: TensorFlow / Caffe / Caffe2 / ONNX / MXNet...
Heterogeneous Runtime
CPU | GPU | APU2.0
Our mission is to be a change catalyst, empowering our partners with smart technology solutions that will inspire them to connect with “next billion” people.

By building technologies that help connect individuals to the world around them, we are enabling people to expand their horizons and more easily achieve their goals.

We believe anyone can achieve something amazing. And we believe they can do it every single day. We call this idea Everyday Genius and everything we do is dedicated to making it possible.
Wrap Up

- Cloud and edge collaboration will be the trend
- Key technologies to tinyML
  - Ultra-low power HW
  - Ultra-light weight runtime
  - More accuracy-efficient model
- MediaTek AI Platform
  - i500/i350/i300 with NeuroPilot
  - MT3620 with NeuroPilot-Micro
- IoT, 5G and AI are the key enabling technologies to our intelligent future
For further information, you may contact:
max.wu@mediatek.com
rituparna.mandal@mediatek.com

You may also visit our website for NeuroPilot online document
https://neuropilot.mediatek.com/

THANK YOU!
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