Endpoint AI Revolution Driven by Standardized Computing Platform

TinyML Asia

Odin Shen
Principal FAE, Arm
2nd Nov, 2021
Semiconductor IP Business

Sparking the World’s AI Potential

The global leader in the development of open compute technology
- R&D outsourcing for semiconductor companies

Focused on freedom and flexibility to innovate
- Technology reused across multiple applications

With a partnership based culture & business model
- Licensees take advantage of learnings from a uniquely collaborative ecosystem

1,900+
Global licenses, growing by 100+ every year

530
licensees
Industry leaders and high-growth start-ups; chip companies and OEMs

190+bn
Arm-based chips shipped to-date

23.7bn
Arm-based chips shipped in 2020
Machine Learning Was Once a Novelty ..... 

Less than a decade ago, a Convolutional Neural Net (AlexNet) won the ImageNet computer vision challenge, and the Machine Learning explosion began
ML is now Mainstream & Deployed – Cloud and Endpoint
Why has (some) Machine Learning Moved to the Edge?

Train in the Cloud. Infer where the Data is Created
Pervasive AI/ML

77% of devices we presently use feature at least one form of AI (1)

270% growth of enterprises using AI in business since 2015 (2)

(1) TechJury
(2) Gartner
ML Usage Growing In All Industries; Still At the Early Stages

**Infrastructure**
- Radio Control (e.g. beam forming)
- Data Traffic Optimisation

**Consumer Electronics**
- Face Unlock
- Object Detection
- Image Augmentation (beautification, focus adjustment, etc.)
- Voice Recognition
- Super Resolution
- Predictive Text

**Sensors**
- Lidar/Radar Improvement
- Audio Sensing
- Vibration Detection
- Sensor Fusion

*Billions of devices today are running ML on Arm*

*Next Step: Trillions of devices*
Specialized Processing Is the Key to ML Deployment

Specialized processing everywhere
Enabling the Internet of Things – ML at the Endpoint

ML Inference at <1mW
Many Different Types of ML Workload in Endpoint Devices

Audio Anomaly Detection

Voice Assistant (endpoint based)

Video Enhancement

- Scalar/Misc.
- Vector/Custom
- Matrix/Neural Net
What Makes Endpoint AI/ML Challenging?

- **Trends in Neural Networks**
  - Larger models → higher accuracy/functionality
  - Increased static memory footprint
  - Increased dynamic memory footprint
  - Increased operations/inference
  - Novel architectures and operators

- **Endpoint ML Constraints**
  - Power
  - Cost
  - Memory
  - Compute
CMSIS-NN

- Open Source: launched 23 Jan’18
- CMSIS-NN has the equivalent role for Cortex-M CPUs as Compute Library has for Cortex-A and Mali
- But flow is entirely offline, creating a binary targeting M-class platform
- DSP instructions in Cortex-M4, M33, M7 & M55
- Will run on Cortex-M0

CNN Runtime improvement

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>New kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative throughput

Energy efficiency improvement

<table>
<thead>
<tr>
<th></th>
<th>Relative Ops per Joule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td></td>
</tr>
<tr>
<td>Pooling</td>
<td></td>
</tr>
<tr>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

https://blog.tensorflow.org/2021/02/accelerated-inference-on-arm-microcontrollers-with-tensorflow-lite.html
Ethos-U NPU’s for Embedded Systems
ML acceleration designed for Endpoint Inference

- Ubiquitous presence
- NN acceleration in software
- Orders of magnitude increase in NN perf
- Easy integration into existing design
- Neural network acceleration

Common software development environment secures any investment made on software development
Cortex M55 + Ethos U55 for On-Device ML
Arm delivers Specialized Processing for next-gen Edge ML

Typical ML Workload for a Voice Assistant

- Faster responses
- Smaller form-factor
- Improved accuracy

**Speed of Inference**
- Cortex-M7: 1x
- Cortex-M55: 6x
- Cortex-M55 + Ethos-U55: 50x

**Energy Efficiency**
- Cortex-M7: 1x
- Cortex-M55: 7x
- Cortex-M55 + Ethos-U55: 25-75x

Partner Silicon Announcements Now
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

TF: TensorFlow
TFL: TensorFlow Lite
Corstone-300 Reference Design

- Unlock performance and power capabilities of the Cortex-M55 processor.
- Helps you build Secure SoCs quickly –
  - Processors,
  - Security,
  - System IP,
  - Software Stack, and
  - Development Tools
Corstone-300 Boards

FPGA Prototyping Board

- FPGA: Xilinx Kintex Ultrascale KU115 FPGA, 1.451k logic cells
  Support for encrypted FPGA Images and Partial Reconfiguration

- Memory: 8MB BRAM
  4GB DDR4-SDRAM (by default, upgradable to 8GB)
  16GB eMMC
  8MB QSPI Flash

- Debug: JTAG
  10-pin Cortex debug connector
  20-pin Cortex debug and ETM connector
  16-bit Trace Macrocell
  ILA for ChipScope Pro™/Identify™
  CMSIS-DAP support

- Board peripherals: USB2.0 Dual port Host Controller
  10/100Mb Ethernet Controller
  uSD Card slot
  Audio (line in/out and mic)
  QVGA Colour Display & Touch Screen - 8-bit parallel interface
  HDMI 1.2 PHY
  Four Virtual UARTs over USB CONFIG PORT
  Eight user LEDs/switches


Arm Ecosystem FVPs

- Download the FVP model for the Corstone-300 MPS3 based platform
  These Corstone-300 models are aligned with the Arm MPS3 development platform and includes both the Cortex-M55 and the Ethos-U55 processors.

  Download Windows
  Download Linux


Demo

Silicon will be shipping this year.
End-to-end Closed Loop Is **The Key** to Scale out of ML

Silicon technology from Arm is here. The Arm Partnership will leverage cores into solutions.
Standards

• Arm believes strongly in open collaboration and standards wherever possible. Examples:
  
  • System and software standards (supports differentiation)
    – E.g. AMBA, Arm System Ready, CoreSight, Arm GIC, Arm SMMU, ASTC, AFBC
  
  • Secure, interoperable software (available to non-Arm HW)
    – E.g. Arm PSA
  
  • AI performance benchmarks
    – E.g. AIIA, ML Commons, EEMBC, and others
  
  • A new open standard: TOSA
    – Tensor Operator Set Architecture
Software and Tools

- Arm ML SDK provides best-in-class ML performance across all of Arm processor (Cortex and Neoverse CPUs, Mali GPUs and Ethos NPUs)
  - Open-Source SDK supports common frameworks and model formats (including Tflite, TFLiteµ, Android NNAPI, PyTorch and ONNX)
  - Quick integration and a seamless developer experience

- ML in device requires Neural Network model optimisation for device
  - Node pruning, weight clustering, quantization and others
  - Arm has enhanced these widely adopted tools:
    – TensorFlow Model Optimization Toolkit (TF MOT)
    – Arm Development Studio
    – Arm Keil MDK
Ecosystem and Developers - Fundamental To Success

Nurturing the world’s most vibrant and successful ecosystem into the AI era

www.arm.com/ai-catalog

- Nurturing a rich, living network of AI partners/expertise
- Developers: ensure deployment of AI on Arm is quick and easy
- Support developers with know-how, insights and information
- Foster a genuine community for interaction with and around Arm
Arm AI Partnership – Harnessing the AI/ML Revolution

Benefits
- Ubiquitous
- Easy to use
- Vast ecosystem

Scalable Processors
- CPU, GPU, NPU

AI Ecosystem
- Standards
- Physical IP
- Support
- Tools
- Software
Unlock Opportunities
Reduce Risks
Change the World

Learn More:
arm.com/solutions/artificial-intelligence/machine-learning